DOCUMENT RESUME

ED 021 784 24 SP 001 516

By- Mayo, Samuel T.

PRE-SERVICE PREPARATION OF TEACHERS IN EDUCATIONAL MEASUREMENT. FINAL REPORT.

Loyola Univ., Chicago, Ill.

Spons Agency- Office of Education (DHEW), Washington, D.C. Bureau of Research.

Bureau No- BR-5-0807

Pub Date Dec 67

Contract- OEC-4-10-011

Note- 125p.

EDRS Price MF-$0.50 HC-$5.08

Descriptors- ACHIEVEMENT TESTS, CHECK LISTS, *EDUCATIONAL NEEDS, *EDUCATION MAJORS, *MEASUREMENT, PRESERVICE EDUCATION, STATISTICS, *TEACHER EDUCATION CURRICULUM

Identifiers- Measurement Competency Test

Because teacher training programs have put relatively little emphasis on the evaluative role of teachers, a project was conducted to determine what teachers need to know, what beginning teachers do know, and what they later learn about measurement. The Measurement Competency Test, developed through consultation with a national sample of experts, was administered in 1964 to a sample of 2,877 senior education majors in 86 randomly chosen teacher-training institutions. Statistical analysis of the data, with that from the 1966 posttest (N=541), revealed that the test scores were unrelated to the kind, selectivity, or location of the institution; scores were related to teaching field, amount of test and measurement course work, and verbal ability. Major conclusions are that (1) there is general agreement on the importance of some measurement compentencies for teachers, but a strong bias against statistics among some teachers; (2) beginning teachers do not demonstrate a very high level of measurement competency, and they show very small gain two years after graduation. It is recommended that some measurement course work be made compulsory, that all be made more meaningful, and that further research be conducted. Included are an 18-item bibliography, the Measurement Competency Test, statistical tables, and materials used for developing the test and conducting the study. (JS)

BR-5-0807
P.A. 24

FINAL REPORT

Project No. 5-0807 / Contract No. OE 4-10-011

24

Pre-Service Preparation of Teachers In Educational Measurement

December 1967

U. S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education/Bureau of Research

SP001516
ED021784

PRE-SERVICE PREPARATION OF TEACHERS
IN EDUCATIONAL MEASUREMENT

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Project No. 5-0807
Contract No. OE 4-10-011

Samuel T. Mayo

December 1967

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

Loyola University

Chicago, Illinois

SP001516

Contents

Acknowledgments v

Chapter I - Introduction 1

The Problem 1
Relevant Literature 2
Background of NCME Committees 3
Purposes 4

Chapter II - Methodology 6

General Overview of Methodology 6
Definition of Measurement Competency 6
    Preliminary *Checklist* Development 6
    Final *Checklist* Development 7
Development of Measurement Competency Test 8
    Item Writing for Measurement Competency Test 8
    Tryout of Test and Allocation of Items to Forms A and B 9
    Discussion of Subscore Classification 9
    Relation of *Checklist* Statements to *MCT* Items . . . 10
Development of the Senior Questionnaire 10
Sampling in First Testing 14
    Selection of Sample of Cooperating Institutions . . . 14
    Selection of Subsample of Seniors within Institutions 18
Follow-Up of Seniors 20

Chapter III - Results 22

Checklist Results 22
    Quantitative Results 22
    Some Qualitative Results 22
First Testing Results 28
    Senior Questionnaire Results 28
    MCT Total Scores 31
    MCT Reliability 33
    Adequacy of Subscores 33
Relationships between *MCT* and Institutional and Personal Variables 35
    Institutional Variables 35
    Personal Variables 36
    Verbal Intelligence and Intellectualism 38
Follow-Up Results 39
    Item Analysis 44

Chapter IV - Discussion, Conclusions, and Implications 49

Discussion of *Checklist* Results 49
Discussion of First Testing Results 50
Discussion of Follow-Up 52
Conclusions 54
Implications 55

Chapter V - Summary 60

The Problem 60
Methodology 60

Results 61
Conclusions 62
Implications 63

References 65

Appendix A - Tentative Outline of Needed Competence in Measurement for Prospective Teachers 67

Appendix B - Letter Sent to Selected Sample of Measurement Specialists and Educators Requesting Cooperation to Respond to *Checklist* 71

Appendix C - *Checklist* of Measurement Competencies 72

Appendix D - *Checklist* Statements Ranked in Order of Mean Response. 78

Appendix E - Measurement Competency Test - Form A 84

Appendix F - Measurement Competency Test - Form B 91

Appendix G - Questionnaire for Seniors in Teacher-Preparation Programs 98

Appendix H - Letter Sent to Institutions Requesting Cooperation to Participate in Senior Testing 101

Appendix I - Summary of Proposed Research 102

Appendix J - Questionnaire for Sample of Institutions Chosen for Graduating Seniors Sample 103

Appendix K - Memorandum to Testing Coordinators for Sample of Graduating Seniors in Teacher-Preparation 104

Appendix L - Questionnaire for Coordinators of Senior Testing . . . 106

Appendix M - Memorandum to Testing Coordinators 108

Appendix N - Directions for Test Administration 111

Appendix O - Chart for Drawing a Random Sample for Varying Sizes of Graduating Class 113

Appendix P - Address Verification Card 114

Appendix Q - Letter Sent to Seniors Requesting Cooperation in Follow-Up Testing 116

Appendix R - Cover Letter Sent to Seniors with Follow-Up Test . . . 117

Appendix S - Follow-Up Questionnaire 118

ERIC Report Resume 119

List of Figures

Figure 1 - Euler Diagram for Original Samples and Follow-up Subsamples 40

Figure 2 - Euler Diagram for Original Samples and Item Analysis Subsamples 45

List of Tables

Table — Page

1. Ranges of Order Numbers of Items in Various Content Categories of the *Measurement Competency Test* . . . 10

2. Behavior Categories of Form A and B Items 11

3. Relations Among Specific *Checklist* Measurement Competencies and *Measurement Competency Test* . . . 12

4. Frequencies of Institutions in National Population and CRP Sample According to Type of Control and Type of Function 15

5. Chi-Square for Representativeness of CRP Sample for National Population on Institutional Control and Function 17

6. Chi-Square for Representativeness of CRP Sample for National Population on Geographical Distribution . 18

7. Sample Size Data for Follow-Up (Person as Sampling Unit) 21

8. Frequency Distribution of Ratings by Content Categories 23

9. Summary of Educational Background of CRP Sample According to Form A, Form B, and Total 29

10. Raw Score Frequency Distribution and Percentile Norms for *Measurement Competency Test*, Form A and B . . . 32

11. Range of Institution Means on *Measurement Competency Test* 34

12. Reliability of *Measurement Competency Test* 34

13. Summary of F-Tests of Significance for Institutional Variables and *MCT* 36

14. Summary of F-Tests of Significance for Personal Variables and *MCT* 37

15. Intercorrelations of *MCT* and Other Variables for Form A and B 41

16. Means and Standard Deviations of Gains for Amount of Tests and Measurement Subgroups 43

17. Item Analysis Data of *MCT* Subsamples Split into Criterion Groups at Median 46

18. Item Difficulties for the *MCT* Follow-Up Subsamples, Pre-Test and Post-Test (Forms A and B) 48

Acknowledgements

The project reported in this document involved many persons from a variety of positions. The following words are a modest attempt to recognize their efforts which made this study possible.

Appreciation is expressed to the Cooperative Research Program of the U.S. Office of Education for providing the necessary funds. Without such support, the work could not have been accomplished.

Recognition is hereby made of the contributions of the Committee on Pre-Service Preparation of Teachers in Measurement of the National Council on Measurement in Education. The NCME committee had been active for several years previous to the initiation of the project. They also served as the Advisory Committee to the project and during the grant period convened annually at the annual NCME-AERA meetings to review progress and to make recommendations. Members of the Committee were: Neal B. Andregg, Howard A. Bowman, Desmond L. Cook, Glen R. Hastings, Irvin J. Lehmann, Samuel T. Mayo (Principal Investigator of Project), Victor H. Noll (Chairman of Committee), John E. Stecklein and Willard G. Warrington.

Several staff persons who worked on various phases of the project should be recognized. Among those who assisted during the initial phase of defining measurement competencies and developing the objective test were Guy Mahan, Harold Messinides and Herbert Paske. Anne Kennard and Frank Trankina did most of the analysis of results from the first testing. In the analysis of follow-up data and final report writing Raynard Dooley and Ronald Bohatch assisted materially.

Contributions were made by several persons outside the project. Item analysis of initial test data and a factor analysis of items were carried out at Ohio State University under the direction of Daniel Stufflebeam, who also advised on experimental design. Max Engelhart and Henry Moughanian aided in writing items for the objective test.

Esther Diamond carried out an ancillary study on variables related to institutional differences in measurement competency and relation of intellectual variability to competency within one institution. Anne Kennard completed a doctoral dissertation on student characteristics related to achievement in measurement classes. Other ancillary studies with the objective test were made by Owen Scott at the University of Georgia, Howard Lyman at the University of Cincinnati, and Raynard Dooley at Northern Michigan University. The *Checklist* was adapted to a survey of about 500 English teachers in Illinois by J. N. Hook and his associates at the University of Illinois (Urbana).

Appreciation is expressed to the many hundreds of persons who responded to the paper-and-pencil instruments in the project, the experts who completed a checklist and the graduating seniors who took an objective test.

To the many other persons, Loyola University faculty and staff, clerical help, who contributed and whom space does not permit mentioning, grateful thanks are herewith extended.

While acknowledging the indispensable assistance from the many people cited above the Principal Investigator accepts full responsibility for this report.

Chapter I

Introduction

The Problem

It is widely recognized that the instructional task of the teacher consists of four steps: (1) Stating objectives in terms of the desired changes in behavior; (2) Choosing materials and methods to bring about the behavioral changes; (3) Providing the actual instructional situation leading to opportunities for learning; and (4) Evaluating the outcomes or behavioral changes, in relation to achieving the original objectives of instruction. Most attention toward improvement of teacher education has been directed to the first three steps. The fourth has been neglected in some respects.

Clearly, measurement and evaluation are essentials of good teaching. Every teacher must make judgments, measure, appraise, and report. He must know how to select appropriately between commercial evaluation instruments, when available, and how to construct his own when appropriate ones are not already available. Furthermore, the teacher must know how to analyze and interpret test scores and how to apply these results in making practical decisions for future courses of action, such as promoting, screening, counseling, etc. No teacher can function effectively without the rudiments of competence in the evaluation matters above. It is recognized, however, that while there are basic measurement competencies required by all teachers, some competencies may be specific to particular grade levels or teaching fields.

Since measurement competency is such a crucial aspect of teaching ability, it follows, therefore, that programs for the preparation of teachers should provide some opportunity to acquire measurement competence. Unfortunately, it is a fact that pre-service programs in teacher preparation, by and large, do not begin to adequately provide for an acceptable set of competencies, whatever criterion one wishes to use. There is ample evidence for this, and some of this evidence is reported in the next section, Relevant Literature.

Relevant Literature

Very few studies have been done, or papers written in the area of the pre-service preparation of teachers in measurement. Of these, only one shows any great similarity to the present project, that one being Robert Ebel's development of an objective test of measurement competency under the auspices of the National Council on Measurement in Education. By and large, the studies have been concerned with the number and type of course offerings in teacher training institutions and certification requirements in measurement of the states. The more important of such studies are described below.

Noll (1955) surveyed requirements of measurement courses for certification in the various states and the coursework offered in measurement in eighty selected teacher-training institutions of four types: large public, large private, state teachers' and liberal arts colleges. He found that 83 per cent offered an introductory course in measurement. Of these, however, only 14 per cent required such a course of undergraduates preparing for certain types of certificates. Only 10 per cent of the states specified a course in measurement for certification, and it was even rare that states recommended such a course as an elective.

Under the auspices of the Committee on Test Utilization of the National Council on Measurement in Education, Allen (1956) surveyed measurement course offerings and opinions relative thereto in 288 teacher-training institutions, obtaining results similar to Noll's. She found also that a majority of the institutions had reference libraries of standardized tests and reported adequate assistance from test publishers. There was less consensus as to the adequacy of instructional materials and methods, and some specific suggestions for improving these were cited from questionnaire responses.

The studies of Noll and Allen are in agreement in showing that an introductory course in measurement is not generally required by state departments of education for a teaching certificate. Most institutions offer an introductory course in measurement, but comparatively few require it for a teaching certificate.

Studies by Davis (1940) and Byram (1933) were in virtually complete agreement in showing that a large proportion of the problems in their work which teachers judge most serious are in the area of measurement and evaluation. Davis reported on 1,075 public school teachers in Colorado while Byram reported on 485 young college teachers.

Noll (1961a, 1961b) reported a study in which he asked seventy-seven seniors in a large midwestern university who were just completing their program of teacher preparation some questions on fundamental concepts and procedures in measurement and evaluation. He also asked the same questions of 108 experienced teachers in summer session at a large eastern university. The answers obtained to the questions showed a serious lack of understanding of the basic concepts and procedures. In the same reference, Noll reported an increase over a seven year period in the number of states requiring a course in measurement for various specific kinds of certificates.

Ebel (1960) described some tests of competence which he developed on an experimental basis. His work on the Committee on the Development of a Test of the Measurement of Competencies of Classroom Teachers has culminated in the production of a set of 250 tested items suitable for inclusion in a test of measurement competence for teachers.

From the above references two conclusions were clear: (1) There was a dearth of systematic and effective preparation of teachers in measurement; and (2) In-service teachers felt strongly their need for competency in measurement and evaluation.

Background of NCME Committees

This project was a continuation of work begun by the Committee on Pre-Service Preparation of Teachers in Measurement of the National Council on Measurement in Education (abbreviated NCME). Victor H. Noll, Professor Emeritus at Michigan State University, was Chairman of this Committee. The Council, since its founding in 1937, has concerned itself with the effective and proper use of measurement in the schools. From 1957 to 1963 (when the proposal for the project was submitted) three NCME committees were active in studying the problem of competency in measurement. In addition to the Committee on the Pre-Service Preparation of Teachers in Measurement, the two other committees had been concerned with in-service preparation in measurement and with the development of a test of measurement competency. Although considerable progress has been made by the committees, all the members were part-time volunteers without funds for the work of the committees. Further work could not have been carried on without funds from a federal agency.

When the project was funded the members of the Committee continued to serve as an Advisory Committee. The names of the members were: Neal B. Andregg, Howard A. Bowman, Desmond L. Cook, Glen R. Hastings, Irvin J. Lehmann, Samuel T. Mayo (Project Director), Victor H. Noll (Chairman of Committee), John E. Stecklein, and Willard G. Warrington.

Purposes

Broadly speaking, the purposes of the project were to determine what teachers need to know about measurement, what beginning teachers actually know at time of graduation, and what they know two years after graduation. More specifically, the purposes were six in number as follows:

1. To develop a clear, practical definition of measurement competencies needed by teachers in general, and also in different grade levels and teaching fields.

2. To obtain reactions to, or evaluations of, measurement competencies by various groups and to study the differences found with a view to discerning the rationale for such differences.

3. To develop an instrument which would provide a valid, reliable measure of the desired measurement competencies. This instrument would be used for administration to a sample of graduating seniors in teacher-training institutions on two different occasions:

(a) immediately prior to graduation; and
(b) two years after graduation.

4. To collect data about undergraduate programs, such as course work taken, curriculum followed, etc., which would be related to measurement competency found at graduation.

5. To relate changes in measurement competency during the two year period to certain variables, such as (a) teaching experience; (b) in-service programs; and (c) graduate study.

6. To interpret findings of the investigation in relation to current programs for preparation of teachers with implications for modification.

CHAPTER II

Methodology

General Overview of Methodology

The project began with the development of the *Checklist of Measurement Competencies* from an existing subject matter outline which had been developed by the NCME Committee on Pre-Service Preparation of Teachers in Measurement Competency. (See exhibit of outline in Appendix C). The *Checklist* was then submitted to a national sample of experts. On the basis of the experts' expressed judgments of the importance of the seventy checklist behaviors, a table of specifications was prepared for developing the objective tests. A tryout form of 150 objective items was used to construct two forms of sixty items each of the Measurement Competency Test.

Definition of Measurement Competency

Preliminary Checklist Development. At the outset it was determined to cast the *Checklist of Measurement Competencies* in terms of expected behaviors on the part of teachers. The *Tentative Outline of Needed Competence in Measurement of Prospective Teachers* was largely a subject-matter outline, although there were some behaviors given. The four-heading format of the *Outline* was preserved in the organization of the *Checklist* and later in the *Measurement Competency Test*. These four headings were (1) Standardized Tests, (2) Construction and Evaluation of Classroom Tests, (3) Uses of Measurement and Evaluation, and (4) Statistical Concepts.

The outline was comprehensive in its coverage of topics in tests and measurements. It reflected the wide gamut of topics to be found in a set of typical introductory textbooks in tests and measurements. Initially the project staff approached the task without preconceived notions whether the *Outline* included the same set of content which the *Checklist* ought to include.

It was soon evident that some topics on the *Outline* would be more important to a teacher than others in terms of emphasis in the teacher's own work. Some general topics seemed to be more the concern of educational specialists or highly experienced teachers than of the beginning teacher toward whom the study was aimed. Therefore, in preparation of the preliminary draft or subsequent drafts of the *Checklist*, the following topics from the *Outline* were omitted: test security, ratings, sociograms, anecdotal records, observations, cumulative records, counseling and guidance, identification and study of exceptional children, curriculum study and revision, and improvement of staff.

At one time the *Checklist* consisted of 120 statements. A revised *Checklist* of ninety-six statements was administered to a local sample of fifty educators whose comments were helpful in producing the final form with its seventy statements.

Final Checklist Development. The final seventy-item form of the *Checklist of Measurement Competencies* (shown in Appendix C) was administered to what were called "experts." These were a purposive sample of measurement specialists and educators. Lists of names of persons considered competent to judge what beginning teachers ought to know about measurement were elicited from the Advisory Committee. In addition, names were selected from membership lists of the National Council on Measurement in Education, the U. S. Office of Education Directory, and the Divisions on Evaluation and Measurement and on Educational Psychology of the American Psychological Association. An attempt was made to represent different types of personnel (such as experienced elementary and high school teachers; school principals and superintendents; college teachers of measurement; measurement specialists in local, state, and private agencies; and guidance workers).

The final mailing list to whom the *Checklist* was sent consisted of 260 persons. They were classified into five groups: teachers, principals and superintendents, college professors, measurement specialists, and miscellaneous (a group considerably smaller than the others, primarily of counselors and school psychologists).

Of the 260 persons canvassed, the final number of usable returns was 185, or 71 per cent, for the five groups combined.

Development of Measurement Competency Test

Item Writing for Measurement Competency Test. In order to determine the competencies in measurement which prospective teachers actually possess, as well as to measure changes in competencies after a two-year period beyond graduation, a comprehensive test was developed for this assessment. It will be recalled that the content categories of competencies in the *Checklist of Measurement Competencies* included:

I. Standardized Tests
II. Construction and Evaluation of Classroom Tests
III. Uses of Measurement and Evaluation
IV. Statistical Concepts

Each statement on the *Checklist* was classified under one of these four content categories.

The ratings of relative importance of *Checklist* content and behavior guided the allocation of *Measurement Competency Test* items to the four categories. In addition, the percentage of test items dealing with specific objectives within each category was also determined, in part, by the ratings of relative importance of *Checklist* responses.

The test items were written by using several kinds of resource material. Sources were: *Multiple-Choice Items for a Test of Teacher Competence in Educational Measurement*, a set of specimen items prepared and arranged by a Committee of the National Council on Measurement in Education under the chairmanship of Robert L. Ebel (1962); the first sixty items of the *Test of Knowledge and Interpretation of Tests (KIT)*, an objective test used in Cooperative Research Project #509 and authored by J. Thomas Hastings (1960); the instructor's manual to accompany Victor H. Noll's *Introduction to Educational Measurement (*1959*)*; the teacher's manual for *Measurement and Evaluation in Psychology and Education (2nd ed.)* by Robert L. Thorndike and Elizabeth

Hagan (1961); and a pool of miscellaneous items from colleagues.

Tryout of Test and Allocation of Items to Forms A and B. Form X, the item analysis tryout form of the *Measurement Competency Test,* consisted of 150 items. Although it had been hoped that an item analysis of a composite of several institutions could be done, problems of scheduling did not permit this. The analysis was therefore based on available data from one large teacher-training institution. Tetrachoric *r* was calculated as the discrimination index and the items which met the statistical requirement of a range of .20 to .70 difficulty index and a .30 validity index were sorted for inclusion in the final form of the test.

With this statistical requirement, 120 items were included in the test to cover the required content and with the view of allowing one minute per item in a two-hour testing period.

The institutions which were to be part of the sample, however, indicated that extreme difficulty would arise from the proposed length of the test. On the advice of the Advisory Committee of the project, two parallel forms of sixty items each were prepared. This permitted one hour of administration time for each form. The planned sample size was doubled and each institution received either Form A or Form B exclusively. Form A is reproduced as Appendix E of this report and Form B is reproduced as Appendix F.

Discussion of Subscore Classification. Very similar content classifications were used for the forms of the test. A distinction in format, however, was the reversal of the ordering of the content areas. As shown in Table 1, Form A began with the *Standardized Test* section, while Form B began with the *Statistical Concepts* section and followed the reverse order. There were four non-overlapping sets of content areas with fifteen items per set for each of the two forms. Table 2 indicates the items for each form of the test, classified into the Knowledge and Application categories.

Table 1.--Ranges of Order Numbers of Items in Various Content Categories of the Measurement Competency Test

CONTENT CATEGORIES	Item Order Numbers	
	Form A	Form B
I. Standardized Tests	1-15	46-60
II. Construction and Evaluation of Classroom Tests	16-30	31-45
III. Uses of Measurement and Evaluation	31-45	16-30
IV. Statistical Concepts	46-60	1-15

Relation of Checklist Statements to MCT Items. Table 3 indicates each item of both forms of the test classified according to the specific competency that was measured in both the content and behavior categories. The table also includes the Checklist statements dealing with each of the four content areas. The reader should bear in mind, however, that there is not a one-to-one correspondence between *Checklist* and *Measurement Competency Test*. Each test item is shown for only one corresponding *Checklist* item, while in reality some test items overlap two or more *Checklist* items as may frequently be the case in test construction.

Development of the Senior Questionnaire

With the intention of relating undergraduate coursework and background variables to test data, a questionnaire was developed to gather the pertinent information. This questionnaire is reproduced as Appendix G . In addition to the identifying information, the organismic

Table 2.--Behavior Categories of Form A and B Items

Behavior Category	Item Numbers					
	Form A			Form B		
Knowledge	2	16	37	1	21	44
	3	17	46	4	22	46
	4	18	47	8	25	49
	5	19	48	9	31	51
	6	20	51	11	32	52
	7	21	53	13	33	53
	8	32	54	14	34	54
	9	33	57	15	35	56
	12	34	58	16	36	58
	13	35	59	17	41	59
Application	1	27	42	2	23	40
	10	28	43	3	24	42
	11	29	44	5	26	43
	14	30	45	6	27	45
	15	31	49	7	28	47
	22	36	50	10	29	48
	23	38	52	12	30	50
	24	39	55	18	37	55
	25	40	56	19	38	57
	26	41	60	20	39	60

Table 3.--Relations Among Specific Measurement Competencies and Measurement Competency Test

Measurement Competencies	Test Item Numbers	
	Form A	Form B
I. Standardized Test (Checklist Statements 1-10)		
Knowledges		
Achievement Test	7, 12	53
Intelligence Tests	2, 9	
Aptitude Tests	6	54
Use of Tests	4	49, 51, 56
Sources of Information	3	52, 59, 46
Familiarities		
Personality Inventory	13	58
Interest Inventory	8	
Projective Techniques	5	
Abilities		
Teacher Made Tests: Contrast	1	
Interpretation of Scores	10, 11	47, 48
Understandings		
Administration of Tests	14	57
Room Conditions	15	
Health Conditions		60
Time Limits		55
General Intelligence vs. Specific Aptitudes		50
II. Construction and Evaluation of Classroom Tests (Checklist Statements 11-23)		
Knowledges		
Teacher-Made Tests	17, 23	34, 41
Item Construction	18, 20	32
Scoring Tests	19, 21	36
Reporting to Parents	16	
Marking Procedures	22	31, 33
Familiarities		
Chart of Content and Behavior	24	35
Item Construction		44
Abilities		
Educational Objectives	25, 26	38, 39
Item Construction	27, 28, 29, 30	37, 40
Understandings		
Correction for Guessing	(none)	(none)
Item Construction		45

III. Uses of Measurement and Evaluation (Checklist Statements 24-36)		
Knowledges		
Validity	31, 32	16, 22
Reliability	33, 34, 35	19, 25
Item Analysis	(none)	(none)
Interpretation of Scores	36	
Familiarities		
I.Q. Range of Ability	37	
Frequency Distribution		21
Abilities		
Diagnostic Test Results	38, 42	26
C.A., M.A., I.Q., and Deviation I.Q.	29	23
Comparison of Two Sets of Data	40	24
Item Analysis	44, 45	20
Understandings		
Percentages	43	28
National Norms	41	30
Standard Error of Measurement		27
Interpretation		29
IV. Statistical Concepts (Checklist Statements 37-70)		
Knowledges		
Mean, Median, Mode	46, 47, 48, 52	13, 14
Comparison of Percentile Rank Scores	(none)	(none)
Ideal of Normal Distribution	58	
Application of Standard Scores	51, 53	11, 15
Non-Normal Distribution		8
Pearson Product Moment Correlation Coefficient	59	9
Familiarities		
Ranking of Scores	57	
Scatter Diagrams		4
Use of Derived Scores	54	1
Graphs	(none)	(none)
Abilities		
Class Intervals		6
Computation of Mean, Median and Mode	55	12
Computation of Semi-Interquartile Range	(none)	(none)
Conversion of Raw Scores to z-Scores	(none)	(none)
Interpretation of Stanines		7
Understandings		
Standard Error of Measurement	60	
Histogram & Frequency Polygon		3
Measures of Variability	49, 50	2
Interpretation		5, 10

variables of age and sex were included, as well as academic background in high school and college.

Information regarding high school background included the number of years of mathematics coursework and science coursework taken. The information regarding college background included the amount of mathematics, science, psychology, and professional education courses. As can be seen from questionnaire statements 17 through 19, special emphasis was given to coursework taken in statistics, and tests and measurements. Other items included the level of teacher preparation and the major and minor teaching fields, as well as student teaching, teaching experience, and transfer pattern.

Sampling in First Testing

Selection of Sample of Cooperating Institutions.

An attempt was made prior to actual testing to secure a representative sample of graduating seniors in teacher-training programs. The sample was obtained by using a fixed-interval design followed by subsampling within institutions. The most complete listing of teacher-training institutions in publication at the time of this phase of the research was *A Manual on Certification Requirements for School Personnel in the United States* by W. Earl Armstrong and T. M. Stinnett (1962). This listing contains the names of 1,061 teacher-education institutions, exclusive of technical schools and junior colleges. Institutions are classified by Function and Control; Type of Control having the Categories of Public and Private, and Type of Function having the Categories of either Teacher-Training Primarily, Universities, and Liberal Arts and General Colleges.

As illustrated in Table 4, of fifty-five Teachers' Colleges, forty-four are Public and eleven are Private. Of 244 Universities, 105 are Public and 139 are Private. Of 762 General Colleges and Liberal Arts Colleges, 190 are Public General and 572 are Private Liberal Art Colleges. The table also includes the frequencies of the institutions in the CRP (Cooperative Research Project) sample for the various categories.

Table 4.--Frequencies of Institutions in National Population and CRP Sample According to Type of Control and Type of Function

Type of Function	Type of Control			
	Public		Private	
	Nat. Pop.	CRP Sample	Nat. Pop.	CRP Sample
Teachers Colleges	44	7	11	4
Universities	105	9	139	10
General and Liberal Arts Colleges	190	12	572	44
Total	339	28	722	58

In the fixed-interval stage of sampling, every n-th institution was identified in a frame constructed from the list of institutions in the *Manual on Certification Requirements*. By setting the size of n at 10, one out of every ten institutions would be chosen and would lead one to expect that at least one out of the eleven private teachers colleges would be chosen by random sampling. To eliminate bias, the institution in the first group of ten institutions was chosen by means of a table of random numbers and all subsequent sampling units were chosen systematically, ten institutions apart in the listing, the sample being proportional and based upon the current Armstrong and Stinnett listing.

After the selection of a sample of slightly more than 200 institutions for administration of the two forms of the *Measurement Competency Test and Senior Questionnaire*, a letter, soliciting cooperation, was mailed to each institution in the sample. This letter is reproduced in Appendix H of the present report. Also included was a *Summary of Proposed Research*, Appendix I . Based upon replies to a questionnaire, Appendix J , that was included with this material, approximately 100 institutions agreed to cooperate. Of these 100 institutions, eighty-six were in the final group who cooperated in testing--forty-four for Form A of the test and forty-two for Form B.

Chi-square tests were run to determine the representativeness of the CRP sample in terms of the variables Type of Control, Type of Function, Combined Function and Control, and Geographical Distribution. Chi-square for observed and expected frequencies of Public vs. Private Institutions was not significant with $\chi^2 = .014$ and df = 1. Chi-square for observed and expected frequencies of Teachers Colleges, Universities, and General and Liberal Arts Colleges was found to be significant at the .01 level, with $\chi^2 = 10.21$ and df = 2. In view of this result, chi-square was run for the combined variable of Type of Control and Function. These results are summarized in Table 5.

Table 5.--Chi-Square for Representativeness of CRP Sample to National Population by Institutional Control and Function

	Type of Control and Function						
	1	2	3	4	5	6	Total
Observed f	7.00 (8.14%)	4.00 (4.65)	9.00 (10.46)	10.00 (11.63)	12.00 (13.95)	44.00 (51.16)	86.00 99.99
Expected f	3.57 (4.15%)	.89 (1.04)	8.51 (9.90)	11.27 (13.10)	15.39 (17.91)	46.36 (53.91)	85.99 100.01
Cell χ^2	3.29	10.86	.03	.14	.75	.12	15.19

1 Public Teachers Colleges
2 Private Teachers Colleges
3 Public Universities
4 Private Universities
5 Public General Colleges
6 Private Liberal Arts Colleges

$\chi^2 =$ 15.19
df = 5
$P <$ .01

The results show significance at less than the .01 level of probability and seem to be due to the heavy weighting contributed by the Private Teachers Colleges to the total χ^2 value. Although the expected frequency in this case is one such college, four were included in the sample on the basis of the sampling plan.

As further analysis to test for representativeness of the sample, a chi-square test was run for Geographical Distribution. These results, as summarized in Table 6, were not significant.

Table 6.--Chi-Square for Representativeness of CRP Sample to National Population by Geographical Distribution

	Geographical Distribution				
	1	2	3	4	Total
Observed f	23.00 (26.74%)	22.00 (25.58)	32.00 (37.21)	9.00 (10.47)	86.00 100.00
Expected F	21.96 (25.54%)	22.29 (25.92)	28.21 (32.80)	13.54 (15.74)	86.00 100.00
Cell χ^2	.05	.00	.51	1.52	2.08

1 Northeast 3 Midwest
2 Southeast 4 West

$\chi^2 = 2.08$
df = 3
P - not significant

Selection of Subsample of Seniors within Institutions. It had originally been hoped that a 40 per cent random sample of all last term (May, 1964) seniors in teacher education programs could be selected by each of the institutions. For institutions with a graduating class of thirty or less, a 100 per cent sample was taken rather than a subsample. This decision was made as a hedge in order to forestall the bias from small samples. Only a minority of the institutions with graduating classes larger than thirty were able to draw their subsample on a random basis. The departure from the original plan arose from inconveniences or hardships which would have occurred from following the random-sampling plan. Some institutions said that they could not compel students randomly selected to participate in the testing. A large proportion of institutions were unable to draw the size of the subsample which had been proposed, resulting in considerable variation away from the 40 per cent figure.

Information concerning the type of sampling that could be carried out was obtained by sending a Memorandum and Questionnaire, Appendixes K and L , to the coordinators of senior testing. The various approaches to sampling, other than random, included testing of volunteer groups, testing nearly 40 per cent of intact groups, nearly 100 per cent of groups, and biased sampling due to lack of compulsory testing. The different forms of the test were randomly allocated to the institutions of the sample. One form was administered exclusively within each institution. Analysis of variance done at a later time on the institutional mean scores by type of sampling, showed no significance for the different types of sampling procedures.

The total number of seniors sought for testing on both forms was approximately 3,000. When answer sheets and questionnaires were scanned for missing data, resulting in elimination of 3 per cent and 4 per cent of the cases, there remained a final sample of 2,877 seniors for both forms. This was a subsample from approximately 7,769 graduating seniors in teacher education at the eighty-six institutions and represented 37 per cent of the group. There were 1,780 seniors who took Form A of the test and 1,097 who took Form B. The percent of students sampled from each geographical area closely approximated the percent of schools from the same geographical region.

Data for the investigation of alternate-form reliability were collected in three additional institutions not included in the sample of eighty-six institutions. The results of these reliability studies are reported in a later section of this chapter.

After the selection of the sample, the test coordinators received test booklets, IBM answer sheets, student questionnaires, *Directions for Test Administration* (Appendix N), and a *Chart for Drawing a*

Random Sample for Varying Sizes of Graduating Class (Appendix O) if the institutions were able to follow the 40 per cent random sampling plan. A *Memorandum to Testing Coordinators* that was included with this material is reproduced as Appendix M of the present report. Testing coordinators were asked to report any difficulties encountered which might have affected the validity of the testing.

Follow-up of Seniors

The 2,877 students of the original sample were followed up in 1965, one year after the original testing. At this time an attempt was made to verify the mailing addresses of the entire sample. Table 7 shows that in this address verification, 1,254 replies were received. During 1966, two years after the original testing, a short preliminary questionnaire was sent to the 1,254 persons enlisting their cooperation in taking the test a second time. From this mailing, 753 affirmative answers were received. During the spring and summer of 1966, copies of the same form the students had taken the first time, along with a short questionnaire on the intervening experiences of the two years were mailed to each of the 753 students who had agreed to cooperate. The final sample of students who returned the completed tests and questionnaires was 541. Of these, the number of A's was 341 and the number of B's was 200.

The purpose of the follow-up test was to determine how much change and what kind of change in measurement competency had taken place among the seniors over the two-year period, and, to relate changes in competency during the two-year period to certain intervening variables, such as teaching experience, in-service programs, and graduate study. The null hypothesis that no gain had taken place during the two years, was postulated. Further null hypotheses were postulated about the relations between intervening variables and gain.

Table 7.--Sample Size Data for Follow-up (Person as Sampling Unit)

Sample	Test Form A	Test Form B	Total
Original Senior (8% of Institutions & 37% of Seniors within Institutions)	1780	1097	2877
Address Verification (One Year Later)	768	486	1254
Agreement-to-Participate (Two Years Later)	465	288	753
Final Follow-Up Participants (Two Years Later)	341	200	541

Chapter III

Results

The results of analysis of data in this chapter can logically be divided into three parts as follows: Checklist Results, First Testing Results, and Follow-Up Results.

Checklist Results

It will be recalled from Chapter II that the *Checklist of Measurement Competencies* was administered to five groups of "experts," namely teachers, principals and superintendents, college and university professors, testing and research specialists, and a miscellaneous group. On the completed Checklists for the 185 usable cases, there were frequent write-in comments. Results from qualitative responses are given, following the quantitative responses below.

Quantitative Results. Means of the responses of experts to each of seventy statements ranged from 1.42 to 2.89 on the three-point scale used in the Checklist. The statements are shown in rank order in Appendix D which indicates that the experts feel that a majority of the competencies on the Checklist are important. Only two of the seventy statements (#9 and #47, which are the first two on the table) showed a majority of responses for the option "Of Little Importance." Other than these, the remainder of the competencies were thought to be "Desirable" or "Essential" by some large proportion of the total group of experts.

Table 8 shows the distribution of ratings of "High" "Medium," or "Low" for the four content categories. Statistical Concepts were thought to be the least important as indicated by the fact that only two of the thirty-four statistical statements were judged as "High." Most of the low ratings for statistics were assigned by teachers in contrast to the other four groups.

Some Qualitative Results. Unsolicited write-in comments on the Checklist form showed both agreement and disagreement on importance of competencies. In the following, only the comments which appeared to be

Table 8.--Frequency Distribution of Ratings by Content Categories

Content Category	Ratings* High	Ratings* Medium	Ratings* Low	Ratings* Total	Identifying Item Nos.
I. Standardized Tests	7	1	2	10	1 - 10
II. Construction & Evaluation of Classroom Tests	7	5	1	13	11 - 23
III. Uses of Measurement and Evaluation	4	9		13	24 - 36
IV. Statistical Concepts	2	18	14	34	37 - 70
TOTALS	20	33	17	70	

*Legend for Ratings

Rating	Range of Means
High	2.65 - 2.89
Medium	2.02 - 2.64
Low	1.42 - 1.98

most provocative were cited. The sheer length of material devoted to dissident views should not be interpreted as indicating majority disagreement. Actually the number of extreme dissidents was few.

There was some consensus that teachers are increasingly to be emancipated from the drudgery of test selection, administration and interpretation. As a result, it was felt that teachers need to know very little about tests and measurement. Emancipation comes on the part of specialists in the schools who shoulder the tests and measurement burdens formerly borne by the teachers. (This may be true in some schools, but the number of such schools is probably less than 10 per cent.) In a few cases, the experts spoke to this point. A junior high principal said that his responses were conditioned by the fact that his testing program is delegated to a specialist. A guidance and counseling supervisor mentioned that a teacher should have competency in all but a very few of the activities indicated by the *Checklist* statements. However, this point of view was contingent upon the assumption that teachers have a testing specialist available. He makes a rather strong point that in the absence of such a specialist, a teacher should either have competence approaching that of a specialist or else the uses of tests should be drastically curtailed to avoid the misuse of test results. He cites particularly the case in which persons improperly claim for tests powers they do not have. He goes further to conclude that, "Much of the hue and cry about testing today is the result of misuse of tests by persons not competent to apply the results to the situation in which they find themselves." Along this same vein, a high school principal believes that teachers should recognize the limitations of their own knowledge in tests and measurements and avoid feeling that they had all the answers to the knotty problems of testing and measurement of ability and achievement.

If teachers are actually going to be relieved of most evaluation duties by specialists in the milenium, it may come as a blessing if we are to believe one of our experts who is a professor of psychology. He opined:

> If many colleges of education would drop some of their courses telling students that teachers must be brave, clean, loyal, reverent, etc., and replace them with other courses, I would mark all of the objectives herein as desirable, and far preferable to the mish-mash now taught. I'm afraid this response would be of very little value to your study, however.

This same professor was pessimistic that most beginning or even advanced teachers would ever acquire many objectives on the *Checklist*.

Opinion was divided on the necessity of the statistical objectives. Most of the consensus was that statistics are necessary for the teacher, but, the experts differed on how much was needed, how deep the preparation should be, whether some statistical concepts would be obvious through common sense and experience, or could be learned on the job, whether the statistical objectives of the *Checklist* should be learned in graduate work rather than in undergraduate, or whether some of our concepts are passe and should be replaced by more progressive concepts.

The wide divergence in statistical needs is illustrated by the citations from three experts. A specialist in a city school system in the South thought that at least one course in statistics should be required, and perhaps as a prerequisite to the introductory measurement course. An elementary principal felt that, while a beginning teacher might not have immediate need for certain of the statistical methods, she should have some exposure to them so that with refreshing, they could be put to use later. A letter accompanying the completed *Checklist* from one elementary teacher in Chicago, illustrated an opposite stand from the two foregoing:

> You might wonder why I marked so many X's in the column "Is of Little Importance." In the first place, the beginning teacher has enough to cope with in learning the fundamentals and school procedures in his or her new job. He or she should not be expected to be familiar with complex terms

> that he or she will not use. So many of these questions deal with higher statistics and unless one is schooled in such courses, and has a job requiring this knowledge, i.e., teaching on a college level, I see little value in them, except as knowledge, but not necessarily application. I believe in making charts and interpretation of test data as simple as possible so elementary teachers, parents, and children can understand them. This is gratifying. This is what we can use on the elementary level. I have a feeling that my responses to your questionnaire will be disappointing to you.

There was one statement for which the consensus was to eliminate it from the repertory of at least some teachers. This was statement #18, "Understanding and application of correction-for-guessing formula to an objective test." It received a relatively low ranking quantitatively (M=1.85, and its rank, fifty-eighth out of seventy.) A primary teacher said that correction formulas are not necessary at the primary level. The author of a textbook on Tests and Measurements thought correction-for-guessing of no importance, "since the concept upon which it is based is spurious."

Opinion was divided on *Checklist* statement #27, "Ability to interpret a profile of subtest results of standardized tests." One respondent thought, "Faced with the profile, a college graduate could hardly fail to understand it. Preparation seems unnecessary." Another respondent thought that interpreting a profile was just common sense. Obviously, these persons are functioning without the benefit of understanding the fallibility of scores and the standard error of measurement. One principal said to leave profile interpretations to the counselor. This same principal would leave statement #32 to the counselor, or to counselor training, as he put it. This is somewhat puzzling when we discover that #32 read, "Knowledge of concepts of validity, reliability and item analysis." One administrator in a test publishing company would also omit the item analysis part of #32 for the beginning teacher.

There was also some feeling against other traditional concepts as taught in Tests and Measurements. Rather, alternative concepts should be substituted, it was felt. Among the traditional concepts so criticized were the standard deviation, normal curve and standard scores. The normal curve was thought to be too abstract for the undergraduate. Score theory as given in statement #68 was also felt by many to be of minor importance. One seemingly constructive alternative concept was voiced strongly by a number of respondents who would emphasize stanines in the preparation of all teachers. One research director in the school system of a midwestern city felt that although many of our statistical objectives could easily be checked as important, he thought it better to select fewer concepts and teach more thoroughly. He would emphasize stanines as the basis for test interpretation. He felt that the concepts in stanines could "be taught quite readily and give a working basis for the use of standard deviation without the student retaining the ability to compute this measure." He went on to say:

> We have been using stanines for interpreting intelligence and achievement tests for the past four years. Principals, counselors, teachers, as well as parents, feel that this is the very finest method of reporting to parents they have seen. Our experience has been that those who have begun to use a simple graph which we have developed, increased their use of this method of test interpretation and are recommending it to others. I have seen so much misuse of test results and lack of understanding that I feel your study has a great deal of possibility. Your request to complete the checklist did not ask for the preceding dissertation, but I feel this is an essential area, and, thought this might provide some basis for interpreting my marking if you care to use it.

Among alternative concepts recommended for inclusion and emphasis in measurement preparation were the deviation IQ and expectancy tables. Furthermore, the concepts of 50 per cent difficulty, "floor," "ceiling," and unimodal symmetry seem to be more useful than the concept of normal distribution.

One or two respondents felt that #5, "Knowledge of sources of information about standardized tests," would be available to any college graduate in libraries. The author does not share this faith in college graduates or librarians and would tend to agree with Thorndike and Hagen in their textbook that although we cannot even make a dent in the specifics of the multitude of standardized tests, we can, at least, show students where to go later for the specific information they need.

Several of the comments as well as correspondence received indicated that the project was already having salutary effects. Some indicated that they could put the *Checklist* objectives to immediate use as a guide for in-service programs for teacher-preparation or for pre-service courses. One principal commenting about the *Checklist* said: "It comprises a beautiful piece of in-service material for a morning workshop in tests with new teachers. I have already so used it. My work is cut out for me with them."

First Testing Results

Senior Questionnaire Results. Tabulation of the student responses to the questionnaire variables revealed that there were 901 men and 1,976 women in the sample. The age range of the students was nineteen through sixty-three with the majority, 2,207, falling in the twenty-one through twenty-three year interval. Table 9 summarizes the characteristics of the sample with respect to educational background.

All but twenty-one students had at least one year of high school mathematics, 2,645 students had two or more years, and over half of the students had three or four years. Although the high school science frequencies are inaccurate due to processing

Table 9.--Summary of Educational Background of CRP Sample
According to Form A, Form B, and Total

High School Mathematics

Years	A	B	Total
0	14	7	21
1	132	79	211
2	531	346	877
3	622	362	984
4	483	301	784

High School Science

Years	A	B	Total
0	33	22	55
1	280	197	477
2	652	356	1008
3	498	309	807
4	518	322	840

College Mathematics

Semester Hours	A	B	Total
0-5	963	617	1580
6-10	542	337	879
11-15	71	33	104
over 15	206	108	314

College Science

Semester Hours	A	B	Total
0-5	94	129	223
6-10	515	365	880
11-15	666	333	999
16-20	233	125	358
over 20	274	143	417

Professional Education

Semester Hours	A	B	Total
0-5	7	12	19
6-10	32	25	57
11-15	134	83	217
16-20	370	235	605
21-25	497	378	875
26-30	198	120	318
over 30	544	242	786

College Psychology

Semester Hours	A	B	Total
0-5	372	171	543
6-10	1078	669	1747
11-15	229	164	393
over 15	103	91	194

College Statistics

Amount	A	B	Total
None	965	637	1602
Part of a Course	603	328	931
One Full Course	179	113	292
More than One Course	35	17	52

Tests and Measurement Courses

Amount	A	B	Total
None	565	301	866
Part of a Course	716	259	975
One Full Course	483	523	1006
More than One Course	18	12	30

When Tests & Measurement Courses Taken

Time	A	B	Total
None	618	315	933
Currently	175	174	349
Last Term	226	214	440
1 Year Ago	567	269	836
2 Years Ago	147	96	243
More than 2 Years	49	27	76

When Student Teaching Taken

Time	A	B	Total
Completed	974	495	1469
Currently	745	529	1274
Not Yet Taken	63	71	134

Continued on next page

Table 9.--Summary of Educational Background of CRP Sample According to Form A, Form B, and Total
(Continued)

Major Teaching Fields in Rank Order

Teaching Fields	A	B	Total
General Elementary	717	322	1039
Social Science	207	159	366
English	205	149	354
Mathematics	124	83	207
Science	120	71	191
Physical Education	109	54	163
Foreign Language	99	56	155
Business and Commercial	81	52	133
Music	41	59	100
Home Economics	36	31	67
Art	12	34	46
Exceptional Children	13	4	17
Industrial Arts, Non-Vocational	12	2	14
Speech Correction	4	5	9
Health Education	0	7	7
Industrial Arts, Vocational	1	4	5
Agriculture	1	3	4
Recreation	0	0	0

Level of Preparation

Level	A	B	Total
Elementary	733	379	1112
Secondary	452	264	716
Both	398	341	739

Where Majority of Work Taken

Institution	A	B	Total
Present Institution	1685	1035	2720
Other	96	60	156
Half-Half	1	0	1

When Transferred

Year	A	B	Total
Freshman	75	37	112
Sophomore	173	111	284
Junior	202	139	341
Senior	38	30	68
Graduate	1	0	1
Did not Transfer	1293	778	2071

Years of Teaching Experience

Years	A	B	Total
None	1687	1024	2711
1	40	27	67
2	17	10	27
3	13	5	18
4	3	2	5
5	8	11	19
Over 5	14	16	30

errors, the results seem to follow the same pattern.

Less than half of the students had taken more than five hours of course work in college mathematics. However, for college science, the majority of students had taken from six to fifteen hours. Psychology ranked between mathematics and science, with the majority of students having taken from six to ten hours of course work. Professional education courses far outweighed the other categories with most students having taken over twenty-one hours of course work in this area. These results seem to be in accord with Conant's (1963) statements concerning the preponderance of education courses required for teacher education. The results for work in college statistics and tests and measurement will be treated more fully in a later section of the present chapter.

The most popular major teaching field was general elementary, with 1,039 students indicating this as their major concentration. Table 9 presents the frequencies of students prepared for each major teaching field in rank order. The sample closely resembled the national population of graduating senior (in teacher preparation) with respect to the percentage of students in the different major field of preparation. The national population figures were obtained from *Teacher Supply and Demand in Public Schools* (1964).

Other background characteristics of the sample, summarized in Table 9, include when student teaching was taken, level of preparation, transfer pattern, and teaching experience. As might be expected of graduating seniors, few had prior teaching experience. The range of years of teaching was from one to twenty for the 166 students who did have prior teaching experience.

MCT Total Scores. For the eighty-six institutions the total number of usable answer sheets for the *MCT* on both forms was 2,877. Of these 1,780 were Form A and 1,097 were Form B. Descriptive statistics on total scores are shown in Table 10 which gives frequency distributions, percentile norms, means and standard deviations. The range of scores for Form A

Table 10.--Raw Score Frequency Distribution and Percentile Norms for *Measurement Competency Test*, Form A & B

	Form A		Form B	
Interval	f	%ile	f	%ile
48-50	12	99+	1	99++
45-47	18	99	1	99+
42-44	41	97	6	99
39-41	90	93	17	98
36-38	140	87	38	96
33-35	225	77	67	91
30-32	276	63	112	83
27-29	280	47	164	70
24-26	266	31	211	53
21-23	196	19	219	34
18-20	135	9	151	17
15-17	53	4	74	7
12-14	37	2	27	2
9-11	6	<1	7	<1
6-8	5	<1	1	<1
3-5			0	<1
0-2			1	<1

	A	B
N =	1780	1097
M =	28.61	24.97
σ =	7.284	6.226

was from six to fifty and for Form B the range was from one to fifty. The two forms of the test did not show a very close parallel. Form B consistently showed itself to be more difficult than Form A. The numerical difference was slightly more than three and a half test score points. Because of this difference and a correlation of only .75 between forms, all subsequent data were analyzed separately by form.

The results of the first testing also indicated considerable variation among the mean scores of the institutions in the sample. We may note from Table 11 that for Form A of the test the institutional means ranged from a low of 20.47 to a high of 35.54. For Form B the means ranged from a low of 17.66 to a high of 34.11. This represents a considerable range for mean scores.

MCT Reliability. Four estimates of the reliability of the *Measurement Competency Test* were made, namely, KR_{20}, KR_{21}, Split-Half, and Alternate-Form-Test-Retest. These results are summarized in Table 12.

The reliability measures were based on data from the total CRP sample, except for Alternate-Form-Test-Retest reliability. The latter was based on ancillary data of five groups of students from three universities outside the sample. Alternate-Form-Coefficients reliability ranged from .59 to .86 with an average of .75.

Adequacy of Subscores. The six *MCT* subscores, previously described in the discussion of Subscore Classification Section of Chapter II, had been set up on an *a priori* basis. All *MCT* answer sheets were scored on the subscores, and their adequacy was studied empirically.

Whenever subscores are set up for a test, there are two potential sources of trouble. First, the few items upon which a subscore is based tend toward too low reliability of the subscore. Secondly, the intercorrelations among the subscores may be so high that they cannot be considered to measure distinct traits. Both of these arose in this project. Since the overall reliability on a total of sixty items

Table 11.--Range of Institution Means on *Measurement Competency Test*

Form	Lowest Mean	Highest Mean	Range	N
A	20.47	35.54	15.07	44
B	17.66	34.11	16.45	42

Table 12.--Reliability of *Measurement Competency Test*

Reliability	Form	
	A	B
KR_{20}	.78	.66
KR_{21}	.75	.60
Split-Half	.78	.68
Alternate Form-Test-Retest	.75	.75

was not relatively high, it was assumed that subscore reliability would be relatively low. The intercorrelations among the six scores were fairly high. A factor analysis of the form content scores showed only one factor. A factor analysis of the sixty items of one form showed no clusters of items and no discernable factor structure.

As a result of the evaluation of subscores above, no further use of subscores was made in the project.

Relationships between MCT and Institutional and Personal Variables

It will be recalled from a previous section of this chapter that there was considerable variance in the total scores of the *MCT* both for individuals and for institution means. The total score distribution of seniors pooled across institutions, in Table 10, showed a heterogeneous distribution. The distribution of institution means showed almost two standard deviations range on the individual norms. An attempt was made to account for this relatively high variance by means of a systematic program of analyses of variance. In the program, two kinds of variables were tested for relationship to test scores. These were (a) institutional variables and (b) student variables. The results are described in the sequel.

Institutional Variables. Using the institution means themselves as scores, a number of variables were tested against the *MCT* by analyses of variance. Institutional variables tested were Control, Type of Institution, Geographical Region, and Selectivity. Results of the tests of significance for both forms are shown in Table 13. There it can be seen that none of the institutional variables showed any significant relation with the *MCT*. The practical result is that the institutional variables do not explain the great variability among institutions.

Table 13.--Summary of F- Tests of Significance for Institutional Variables and *MCT*

Institutional Variable	P Levels	
	Form A	Form B
Control (Public vs. Private)	>.05	>.05
Type of Institution (Teachers College, Liberal Arts, or University)	>.05	>.05
Geographical Region (Northeast, Southeast, Midwest, or West)	>.05	>.05
Percentage of Students Within Institution Taking Tests and Measurements (0-39%, 40-89%, 90-100%)	>.05	>.05
Selectivity (Highly Selective, Very Selective, or Unclassified*)	----	>.05

*Institutions were classified directly from listing in Appendixes section of *Comparative Guide to American Colleges* by James Cass and Max Birnbaum, Harper and Row, 1964.

Personal Variables. In studying personal variables, seniors were first pooled across institutions. Personal variables tested against *MCT* were Sex, Teaching Field, Amount of Tests and Measurements Taken, and, Amount of Statistics Taken. Results of the tests of significance for both forms are shown in Table 14. There it can be seen that Sex was non-significant, while the remaining showed high significance. The practical results are that sex is unrelated to *MCT* score while teaching field and amount of coursework in tests and measurements or statistics are related to *MCT* score. Some comment upon the nature of the relationships is in order.

Table 14.--Summary of F-Tests of Significance for Personal Variables and MCT

Personal Variable	Forms					
	A			B		
	N	MCT Mean	p level	N	MCT Mean	p level
Sex			>.05			>.05
Male	557	28.228		344	25.023	
Female	1223	28.795		750	24.951	
Teaching Field			<.001			<.001
General Elementary	715	28.396		322	23.711	
English	205	27.585		149	26.564	
Mathematics	124	33.177		83	29.289	
Science	119	31.815		71	26.507	
Social Science	208	28.212		158	25.101	
Art	12	25.250		34	23.824	
Music	41	24.512		59	24.068	
Foreign Language	99	28.909		56	24.393	
Business and Commerce	81	29.938		52	25.519	
Industrial Arts (Vocational)	1	26.000		4	25.500	
Industrial Arts (Non-Vocational)	12	24.750		2	23.000	
Agriculture	1	32.000		3	26.667	
Home Eonomics	36	26.000		31	24.710	
Physical Education	109	25.642		54	21.815	
Exceptional Children	13	32.385		4	30.250	
Speech Correction	4	27.750		5	17.000	
(17)	-	-- ---		7	21.143	
Amount of Tests and Measurements			<.001			<.001
More than One Course	18	30.83		12	25.25	
One Full Course	483	30.08		523	25.85	
Part of Another Course	714	30.01		258	25.62	
None	565	25.54		301	22.89	
Amount of Statistics			<.001			<.001
More than One Course	35	35.871		17	28.353	
One Full Course	179	29.760		113	26.708	
Part of Another Course	601	30.556		327	26.000	
None	965	26.953		637	24.049	

At the descriptive level, an interpretation of Teaching Field (restricted to fields with the largest number of cases) showed the following: Mathematics and Science were the highest of any fields on Form A; Mathematics was highest on Form B; Business and Commerce was fairly high on both forms; Social Science, Foreign Languages, and Home Economics were in the middle range on both forms; special subjects like Art, Music, and Physical Education, were low on both forms; English and General Elementary showed inconsistencies. On Form A, significant differences by the *t* test were found between each of the following pairs: Mathematics and Foreign Language, Mathematics and Business and Commerce, and Science and Foreign Language.

When the category means of Amount of Tests and Measurements Taken were examined, it was seen that three of the categories were very close together, while the fourth was very different. Therefore, *t* tests were run between pairs of means. These showed that there were no significant differences among groups with various amounts of Tests and Measurements Taken, but there was a significant difference between scores of students taking no test and measurements and students taking any amount at all.

The pattern of means for Statistics, while similar, showed less uniform results than Tests and Measurements.

The practical result of the latter two analyses was that any amount of coursework in tests and measurements or statistics is associated more often with superiority in measurement competency.

Verbal Intelligence and Intellectualism. Suspecting that general mental ability might account in part for variance on the *MCT*, two kinds of ancillary studies were made.

In the first, a correlation between Miller Analogies test scores, and the *MCT* for one institution, yielded a significant correlation of .56 for 215 cases. An analysis of variance to determine the relation of particular teaching fields to Miller Analogies scores for the same institution also yielded significant results.

In the second study, a correlation between Astin's "Intellectualism" factor and the *MCT* mean scores of students in the participating institutions, yielded significant results for Form B with an r of .46 for thirty-six cases.

In summary, it appears that the variables labelled as "personal" are the best explanation of variance on the *MCT* means for institutions.

Follow-Up Results

From the original samples of pre-test subjects, 341 subjects who originally took Form A, and 200 subjects taking Form B, cooperated two years later to be retested with the same forms of the *MCT*. The pre-test and post-test data from these post-test subsamples and the data from the original samples were used in the following analysis. Figure 1 portrays with an Euler diagram scheme the important data for the various samples and subsamples. Appropriate t tests were calculated between the pre-test and post-test means of the 341 case subsample of Form A; between the pre-test and post-test means of the 200 case subsample of Form B; between the pre-test means of the 341 case and 200 case subsamples of Forms A and B; between the post-test means of the same subsamples; between the means of the original sample and the 341 case pre-test subsample of Form A; and, between the means of the original sample and the 200 case pre-test subsample of Form B. All of these t tests were statistically significant at $p < .0005$.

The data show further that Forms A and B were not strictly comparable, Form B being the more difficult. This difference was still significant after the two year interval. Both form subsamples had gained significantly over the two year interval. There is also some evidence to show that the subjects who cooperated for retesting constituted subsamples which performed significantly better than the original samples. However, in the gain studies, of course, each person served as his own control. The evidence comes from t tests between means of the original and follow-up samples for the forms as shown in Figure 1.

Figure 1.--Euler Diagram for Original Samples and Follow-up Subsamples

MCT Number of Cases, Means, and Standard Deviations for Original Samples and Follow-up Subsamples, Forms A and B

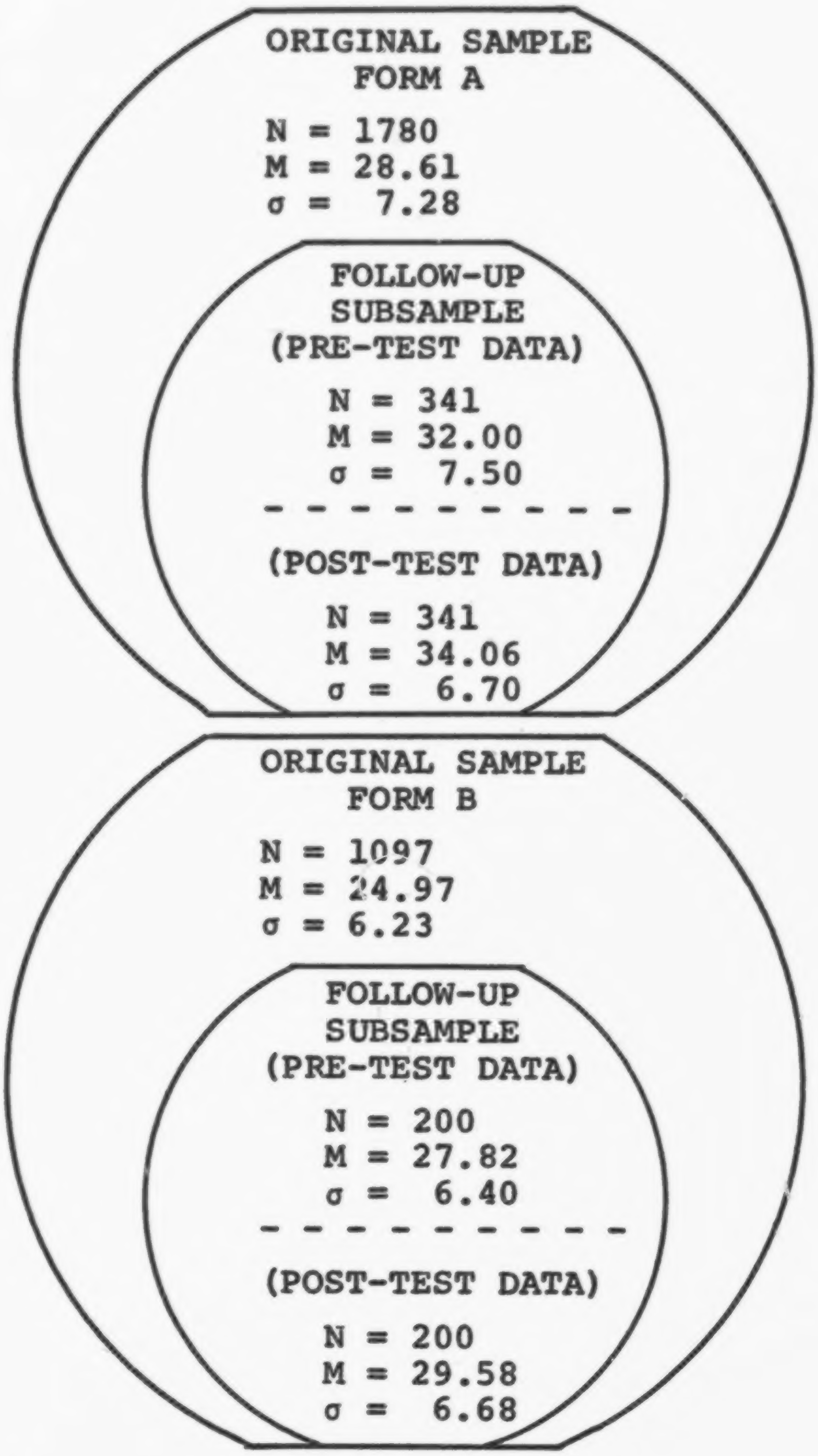

ORIGINAL SAMPLE FORM A
N = 1780
M = 28.61
σ = 7.28
FOLLOW-UP SUBSAMPLE (PRE-TEST DATA)
N = 341
M = 32.00
σ = 7.50
(POST-TEST DATA)
N = 341
M = 34.06
σ = 6.70
ORIGINAL SAMPLE FORM B
N = 1097
M = 24.97
σ = 6.23
FOLLOW-UP SUBSAMPLE (PRE-TEST DATA)
N = 200
M = 27.82
σ = 6.40
(POST-TEST DATA)
N = 200
M = 29.58
σ = 6.68

Table 15.--Intercorrelations of *MCT* and Other Variables for Form A & B*

Form A (Below Diagonal)
Form B (Above Diagonal)

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1.		13	22	62	03	06	-03	10	-40
2.	28		21	-03	-11	03	01	-03	-19
3.	31	23		15	02	03	-15	10	-08
4.	70	16	27		08	06	-13	14	47
5.	02	-07	03	02		41	08	12	07
6.	05	-06	03	00	36		03	13	00
7.	-05	-05	-03	-10	11	06		-66	-11
8.	04	04	01	09	14	-05	-66		05
9.	-51	-18	-09	26	00	-06	-05	05	

1. Score on First Test
2. Amount of T-M Taken
3. Amount of Statistics Taken
4. Score on Second Test
5. Teaching Experience
6. In-Service Training
7. Graduate Study
8. Number of Graduate Semester Hours
9. Gain Score

*Decimal points have been omitted throughout

Results showed that the amount of gain for the A and B groups was slightly more than two test score points. The standard deviations of Forms A and B on original testing were 7.28 and 6.23 points respectively. Thus the average amount of gain across all persons amounted to about one-third of a standard deviation, which is significantly different from zero at the .01 level.

Relations of six variables to gains on the MCT were investigated. Three were pretest variables. These were (1) teaching field, (2) amount of tests and measurements course work taken, and (3) amount of statistics course work taken. The three post-test variables were (1) teaching experience, (2) in-service training, and (3) graduate study.

An analysis of variance indicated that there were no significant differences between the teaching field that the follow up samples had chosen in college. on the size of the gain on the test-retest situation. It had been expected that students in Mathematics and Science would have shown a different amount of gain than those in other teaching fields because of their strong quantitative background and orientation.

The gain score means, the standard deviations, and the number of respondents in each category for the variable, Amount of Tests and Measurement Taken, are reported in Table 16. As the number of Tests and Measurements courses taken increases, the smaller the gain in the re-test situation. In fact, the differences in mean size were significant at .01 level of probability for Form A and at the .05 level for Form B when a one way analysis of variance was performed. Thus an inverse relationship exists between the amount of tests and measurements taken and the gain scores, although it should be recognized that this is an artifact.

For Form A the use of *t* between the means of the none group and part of another course group showed non-significance. Furthermore, differences between one full course and more than one course were non-significant. However, if the first two groups and the last two groups are combined, then there is a

Table 16.--Means and Standard Deviations of Gains for Amount of Tests and Measurements Subgroups

Amount of Tests and Measurements Taken and Gain		Form A (341)	Form B (200)
None	Mean	3.184	2.960
	S.D.	5.428	5.564
	N	87	50
Part of another course	Mean	2.543	2.583
	S.D.	5.709	5.142
	N	140	60
One full course	Mean	.620	.632
	S.D.	5.104	5.878
	N	108	87
More than one course	Mean	.333	2.000
	S.D.	4.955	3.559
	N	6	3

significant difference. We may say then, that people who have had at least one full course showed less gain than those who had had less than one full course. Or to put it another way, the greatest gain was shown by those people who had had less than one full course.

When *t* was calculated for the Form B means, those people who had no coursework in test and measurements showed a significant difference in relation to the other three variables; part of another course, one full course and more than one course. An examination of the score gains for all four variables would indicate that those who have not had any training in tests and measurements were the ones who achieved significantly different gain scores, a result not unlike that found in Form A.

The correlation of the amount of tests and measurements taken with gain scores is -.1822 for Form A and -.1904 for Form B.

The remaining four variables, Amount of Statistics Coursework Taken, Teaching Experience, In-Service Training and Graduate Study, did not indicate a significant difference among their gain scores.

There was a -.09 correlation on Form A between Amount of Statistics Coursework Taken and Gain score. For Form B the correlation was -.08.

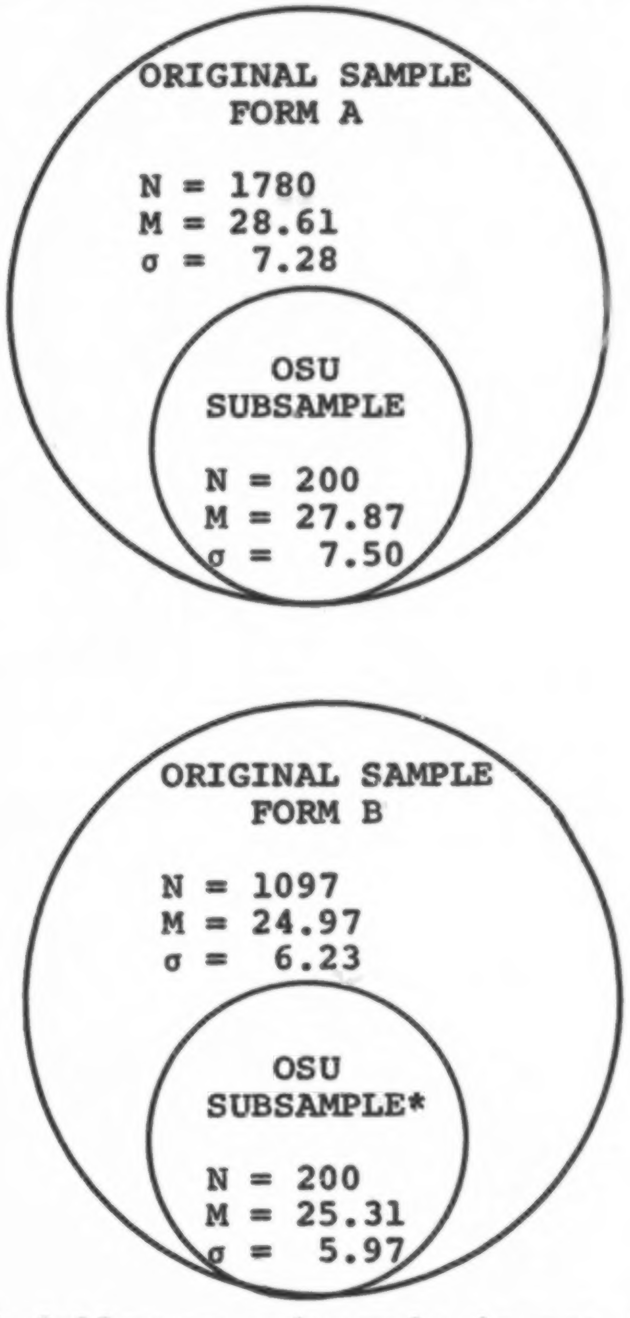
The correlation between the Graduate Study variable and Gain score was .05 on Form A. The correlation was -.11 on Form B.

In summary, five of the six variables showed no relationship to measurement competency gain.

Item Analysis . A comprehensive item analysis was carried out on the first testing seniors. Random subsamples of 200 cases (each described in Figure 2) were taken from each of the two forms. Item analysis was done with the 200 person samples, on a large computer at Ohio State University. Table 17 shows from computer printout, the difficulties, and three kinds of indexes of discrimination for each item. As compared with item analyses of other similar cognitive tests in the author's experience, the *MCT* seemed adequate from an item characteristic viewpoint.

Table 18 shows the difficulty coefficients for each item on each form for the follow-up samples. It will be noted that most of the gains in difficulty are modest and that there are a fair number of negative gains. The few fairly large gains (i.e., positive changes of from .10 or larger) might easily have been due to chance.

Figure 2.--Euler Diagram for Original Samples and Item Analysis Subsamples


ORIGINAL SAMPLE
FORM A
N = 1780
M = 28.61
σ = 7.28
OSU
SUBSAMPLE
N = 200
M = 27.87
σ = 7.50
ORIGINAL SAMPLE
FORM B
N = 1097
M = 24.97
σ = 6.23
OSU
SUBSAMPLE*
N = 200
M = 25.31
σ = 5.97

*The 200 follow-up subsample is not the same as the 200 subsample used for item analysis although there is some overlap.

Table 17.--Item Analysis Data of MCT Subsamples Split into Criterion Groups at Median*

FORM A					FORM B				
Item No.	Diff.	D	ϕ	r (pt.bis.)	Item No.	Diff.	D	ϕ	r (pt.bis.)
1	56	24	24	29	1	25	23	26	30.
2	57	35	35	41	2	25	-2	-3	01
3	50	11	11	16	3	13	9	13	15
4	58	4	4	03	4	43	23	23	31
5	47	28	28	34	5	75	16	18	21
6	39	26	27	31	6	13	10	15	21
7	66	24	26	34	7	35	10	9	08
8	64	28	29	41	8	34	29	30	38
9	63	12	12	22	9	23	12	14	31
10	59	24	24	35	10	16	5	7	06
11	33	22	23	31	11	55	11	11	23
12	61	23	23	30	12	70	23	25	28
13	18	9	11	17	13	56	-4	-4	-01
14	48	25	25	27	14	22	11	14	26
15	34	-5	-5	-12	15	09	8	14	20
16	34	16	18	19	16	60	22	22	28
17	67	25	28	31	17	12	-1	-2	06
18	63	11	11	18	18	33	32	34	32
19	35	19	20	12	19	54	30	30	37
20	42	24	24	26	20	38	22	22	28
21	25	29	33	40	21	37	18	18	22
22	77	18	21	34	22	55	25	25	32
23	24	5	6	15	23	28	12	12	14
24	39	27	28	28	24	24	17	19	29
25	60	27	27	34	25	31	12	13	20
26	32	7	7	11	26	22	8	10	14
27	76	24	27	43	27	58	17	17	27
28	72	27	29	34	28	40	21	21	20
29	51	20	19	26	29	72	15	16	26
30	75	32	36	45	30	40	7	7	19
31	33	15	16	27	31	28	17	19	16
32	50	21	21	30	32	70	20	22	30
33	46	39	39	41	33	56	29	29	25
34	14	00	00	05	34	51	13	13	20
35	30	10	10	09	35	26	-8	-10	-02
36	68	20	21	31	36	58	31	31	38
37	54	17	15	21	37	26	2	2	17
38	22	13	14	11	38	25	3	3	10
39	65	30	30	30	39	39	10	10	16
40	45	15	14	21	40	64	37	39	32
41	59	26	25	36	41	88	15	24	29
42	50	34	32	29	42	47	29	28	31
43	64	39	39	39	43	49	26	26	35
44	67	33	33	41	44	48	21	21	21
45	49	28	26	42	45	47	20	18	09
46	59	22	20	35	46	40	4	3	02
47	29	25	26	36	47	34	17	17	28
48	22	14	16	20	48	16	00	-1	-03
49	42	11	10	24	49	62	15	15	16
50	17	6	7	10	50	33	20	20	27

Continued on next page

Table 17.--Item Analysis Data of MCT Subsamples Split into Criterion Groups at Median* (Continued)

FORM A					FORM B				
Item No.	Diff.	D	ϕ	r (pt.bis.)	Item No.	Diff.	D	ϕ	r (pt.bis.)
51	31	22	23	35	51	70	12	12	21
52	70	17	17	24	52	27	1	1	00
53	37	27	28	31	53	84	17	23	33
54	16	9	12	21	54	37	12	11	26
55	28	4	4	03	55	31	16	17	24
56	38	18	19	27	56	47	23	23	24
57	23	7	7	22	57	27	24	27	37
58	35	14	13	24	58	73	23	25	24
59	42	29	30	33	59	45	34	34	38
60	42	23	24	28	60	50	17	16	19

*Decimal points have been omitted throughout

Table 18.--Item Difficulties for the MCT Follow-Up Sub-samples, Pre-Test and Post-Test (Forms A and B)*

FORM A (341 cases)						FORM B (341 cases)					
ITEM	PRE DIFF	POST DIFF	ITEM	PRE DIFF	POST DIFF	ITEM	PRE DIFF	POST DIFF	ITEM	PRE DIFF	POST DIFF
1	66	68	31	44	40	1	32	42	31	21	30
2	71	78	32	56	60	2	21	24	32	75	78
3	60	56	33	60	65	3	14	14	33	67	68
4	60	62	34	16	13	4	57	54	34	57	50
5	58	64	35	31	26	5	74	79	35	22	22
6	37	41	36	70	78	6	16	18	36	62	72
7	71	76	37	64	69	7	31	32	37	30	31
8	71	76	38	26	29	8	39	50	38	22	22
9	66	71	39	80	81	9	28	26	39	46	50
10	62	76	40	56	56	10	18	18	40	64	80
11	40	32	41	65	76	11	59	58	41	94	95
12	69	76	42	64	67	12	76	82	42	53	66
13	23	26	43	72	77	13	56	54	43	54	60
14	58	57	44	77	86	14	29	30	44	60	64
15	31	30	45	65	73	15	16	20	45	54	56
16	42	52	46	66	73	16	61	70	46	41	45
17	72	78	47	41	44	17	20	18	47	45	49
18	74	76	48	29	28	18	38	48	48	13	14
19	43	38	49	39	52	19	61	62	49	68	74
20	54	55	50	17	24	20	44	40	50	42	39
21	36	40	51	41	42	21	42	38	51	67	67
22	84	86	52	74	82	22	60	56	52	35	26
23	27	22	53	45	51	23	27	26	53	87	91
24	44	56	54	29	27	24	32	30	54	49	58
25	68	72	55	27	26	25	34	32	55	35	44
26	33	36	56	45	47	26	22	30	56	50	56
27	85	90	57	23	28	27	63	72	57	33	45
28	72	86	58	43	44	28	42	48	58	76	82
29	60	62	59	49	57	29	81	80	59	57	55
30	82	84	60	51	63	30	49	64	60	50	54

*All decimal points have been omitted throughout

Chapter IV

Discussion, Conclusions, and Implications

Discussion of Checklist Results

Responses of a selected composite of five groups of experts to the *Checklist of Measurement Competencies* (a checklist of behaviors representing knowledges and skills in tests and measurements) showed agreement with the *Checklist*. (The five groups were teachers, principals and superintendents, college and university professors, measurement and testing specialists, and a miscellaneous group.) The *Checklist* when constructed, had represented a domain of content and behavior common to many textbooks in measurement and, in addition, common to the experience and judgment of specialists in college teaching and infra-college educational staffs.

Results from experts' responses to the *Checklist* showed general agreement on importance of the statements of competencies. This was further strengthened by the qualification that even though a competency was rated low for beginning teachers, it might be essential for an experienced teacher. It is well to ask whether teachers will attain such competencies systematically in graduate work, through in-service training, or through self-study. It was gratifying to find general agreement with the *Checklist* behaviors. Almost all are considered important to teachers at some field or level.

The most striking interaction between kind of expert and kind of competency occurred with teachers and statistics. Teachers rated statistics competencies largely low. Conversely, most of the endorsement as important occurred in the areas of standardized tests, teacher-made tests and, uses of tests. One possible redeeming feature in the teaching of statistics was shown by the sentiment of some college and university professors to play down the importance of the traditional statistical topics, and to play up more enlightened approaches.

There was great diversity of opinion on a number of controversial topics upon which the experts qualified their responses. Among these moot topics were the issues of whether competencies belonged in the undergraduate, graduate, or in-service phases of preparation; whether the teacher would function with or without the services of a specialist in testing; whether formal preparation in statistics was needed and when; and whether some competencies are transferable automatically through formal education and application of intelligence and common sense.

Discussion of First Testing Results

The first testing of the graduating seniors in 1964 provided data upon which to evaluate the test itself and also the status of measurement competencies of the seniors.

In comparison with the usual cognitive tests of comparable type and length, the *MCT* seemed adequate from the standpoint of reliability, discrimination, and item difficulty. The forms, however, lack comparability in many respects. Form B has yielded consistently lower scores. Although attempt was made to produce parallel forms, they did not appear to correlate highly enough to warrant interchangeability. In some of the analyses, they yielded opposite results in tests of significance or in certain trends.

The means for the two forms on the total sample, as shown in Table 5, are both lower than the recommended 50 per cent. More important is the conclusion that on a test constructed so as to subsume content and behaviors judged to be important, the seniors did not distinguish themselves. Table 10 shows some individuals making low scores in the chance region.

It will be recalled that the mean *MCT* scores for the institutions showed great variability, viz. two standard deviations on the basis of student scores pooled for all institutions. Two variables were hypothesized to account for this institutional variability; institutional and personal. Institutional variables hypothesized were Control, Type of Institution, Geographical Region, and Selectivity. None of these showed a significant relation to scores on the

MCT. This was a surprising outcome, since any one of the variables would have been expected to be related, in view of the widespread beliefs that institutions of different kinds in different regions and with differences in recognized prestige and high academic standards, also differ in demonstrated outcomes in achievement. Therefore, the conclusion can be made that the variance in institutions on measurement competency could not be explained on the basis of systematic, *a priori* classifications of institutional characterisctics. One explanation may be suggested. It was, however, untestable in the present study. When the *MCT* was administered to the seniors, none of the project staff were present. Proctors were supplied by the institutions themselves. Very little of the details of conditions under which the test was administered are known. It seems reasonable to speculate that a substantial part of the variance among institutions could have arisen from differences in testing conditions (such as working time, kind of instructions, set and motivational conditions, etc.). Such a variable would tend to be common to all persons within a testing group or within an institution. This is what Prof. E. F. Lindquist has called "Type G Error" in his book, *Design and Analysis of Experiments in Psychology and Education*, Houghton-Mifflin, 1953.

After data were pooled across institutions, the relations among student variables and *MCT* were studied. Sex showed no relation, while amount of tests and measurements taken, amount of statistics taken, and teaching field were related. Any amount of coursework in measurement or statistics resulted in higher achievement on the *MCT*. The pattern of relative achievement in the various teaching fields, while fraught with small score differences and inconsistent results between the two forms, would suggest that the highest *MCT* score would tend to be made by mathematics and science (both "academic" fields) while the lowest *MCT* scores were made in the "special fields" (the non-academic). There may be several explanations. Logically, one would expect more communality between mathematics and science courses and measurement which involves quantitative and applied science orientation. Furthermore, vocational interests of mathematics and science majors would be expected to differ markedly from those in "special fields" and in the direction of measurement. Finally, the obtained correlation found between verbal ability and *MCT* and, between *MCT* and major fields, suggests academic aptitude as a possible explanation.

The fact that verbal intelligence is related substantially to *MCT* raises the question (possibly disturbing to professors of measurement) that the ability to respond correctly to *MCT* items may result in large measure from general intelligence as compared with transfer from specific learning in measurement courses.

Discussion of Follow-Up

The purpose of the follow-up was to determine how much change and what kind of change in measurement competency had taken place among the seniors over the two-year period, and, to relate changes in competency during the two-year period to certain intervening variables, such as: teaching experience, in-service programs, and graduate study. The null hypothesis that no gain had taken place during the two years, was postulated. Further null hypotheses were postulated about relations between intervening variables and gain.

The amount of gain found for both the Form A and Form B groups was slightly more than two test score points. This gain was statistically significant, but in a practical sense, was only one-third of a standard deviation and therefore, small.

The Principal Investigator was able to achieve gains as high as ten raw score points on the *MCT* in his own measurement classes under the conditions of using the *Checklist* and *MCT* as a basis of preparing the topical outline for the course and planning the daily class activities. Perhaps this represents an upper limit of gain as a goal to strive for.

Six variables were tested against gain. These were teaching field, amount of tests and measurements coursework taken, amount of statistics coursework taken, teaching experience, in-service training, and graduate study. Only one was significant statistically, that one being amount of tests and measurements taken. This latter relationship was the inverse of results from the first testing. Specifically, there was a positive relationship between amount of tests and measurements and first testing score, while there is a negative relationship between amount of tests and measurements of gain.

How shall the gain results be explained? There are at least two possible explanations. First, persons who had had little or not tests and measurement had more to learn, whereas the ones who had tests and measurements may have reached a saturation point. Furthermore, the less sophisticated would have less difficult things to learn which had already been mastered by the more sophisticated who were learning more difficult things and showing less gain.

A second explanation is that this difference is due to the regression effect that is always present in the gains type of study. Regression must have taken place here, since the analysis of variance of the pre-test results showed that the people who had had the least tests and measurements made the highest gain scores on the *MCT*. This, of course, is in the direction that would be expected by the theory of gains studies. To put it another way, those persons who had made high scores by chance on the first testing would tend to make lower scores by chance on the second test, while those who made lower scores due to chance on the first testing would tend to make higher scores on the second testing. In both cases, retest scores regress toward the mean.

Five variables showed no relation to gains on the *MCT*. They were teaching field, amount of statistics coursework taken, teaching experience, in-service training, and graduate study. The matter begs for some explanation. It is difficult to explain why the major field and amount of statistics coursework bore no relation to gain in measurement competency. One might have expected teachers of mathematics, science and business, e.g. to have improved more than others by using quantitative concepts and being more conscious of statistics and measurement. The summary relationship can be described only as follows: The kind of undergraduate curriculum which a graduate had taken did nothing to enhance or detract from gain in measurement competency, with the exception of relevant coursework in measurement.

However, an attempt can be made to explain the lack of effect of two-year intervening experiences upon gain in measurement competency. Recall that gain was unrelated to graduate study, teaching experience, and in-service training. Graduates must be increasing their competency in some areas, but it is not in Tests and Measurement as measured by the *MCT*. The explanation must be that their experiences are not relevant to measurement knowledge and skills.

There were some honest differences of opinion on the interpretation of results of the study among the Principal Investigator and members of the Advisory Committee. For example, the gain of slightly more than two *MCT* test score points, which is one-third of a standard deviation, is statistically significant! There is no debate about this! The debate comes in the attempt to decide whether this difference is large in a practical sense. Would it be large enough to make an important distinction in a teacher's behavior when observed on pre and post occasions? The Investigator believes not!

Conclusions

Several conclusions may be drawn from the results of this study, from some ancillary studies related to the project, and from the interchange of ideas with professional colleagues. Following are the conclusions from which the later implications were made:

1. There is general agreement as to the importance of teachers possessing certain "core" competencies in measurement, but there is diversity in thinking about how and when they should be learned.

2. Some teachers, especially elementary teachers, have a strong bias against statistics, apparently because they see no relation to their work.

3. Beginning teachers, as a whole, do not possess, to a high degree, the knowledges and skills in measurement which have been defined as important by measurement experts. The Principal Investigator had assumed that Coursework in Tests and Measurements would be expected to produce an increase in measurement competency which should be measurable on an objective test. Furthermore it was assumed that if a negligible difference were found between test scores from persons exposed to two different treatments, then one would be in a position to conclude logically that there was evidence that whether a person had either one treatment or another one made little practical difference in observed measurement competency. Among the variables which relate to measurement competency at time of graduation are the teaching field, and whether coursework in tests and measurements and statistics were taken. Persons who had taken any amount of statistics or tests and measurements were superior to those who had had none. Persons from teaching fields of mathematics and science showed superiority to those of other teaching fields. Such differences were, however, modest.

4. During a two year period after graduation, graduates of teacher preparation programs show only a small improvement in measurement knowledges and skills. Only the amount of tests and measurements taken showed any relation to gain in measurement competency over the intervening period, and this was in inverse relationship. Variables which showed no relation to gain were amount of statistics, teaching field, teaching experience, in-service training, and graduate study.

5. Verbal ability was significantly related to measurement competency and to teaching field.

Implications

It is evident that the entire set of competencies sampled by the *Checklist* and the *MCT* should not necessarily be expected to be mastered by the beginning teacher. Furthermore, even among experienced teachers, not every teacher would necessarily need every competency in the set. Different subsets of competencies would be needed by elementary as contrasted with secondary teachers. From this viewpoint, the average performance of the seniors on the *MCT* (i.e., between 40 per cent and 50 per cent on a percent of maximum type score) would not be disappointing. Nevertheless, the level of performance is still far from mastery. Perhaps local norms should be developed on a measurement competency test and differentiated for various teaching fields. In this way, perhaps, "quality control" of measurement competency could be assured during training.

Although this project did not provide the evidence to test the supposition, it suggests that perhaps many of the graduates of our teacher-training programs, although learning some measurement competencies, do not become deeply involved in the problems and practices of evaluation and are not sensitive to the need to commit themselves toward raising their level in measurement competency. It is well known, for example, that some teachers habitually construct poor tests without realizing how poor they are, and without knowing first that they should improve, and second, how they can improve. Evidence from the project and from the personal experience of the Principal Investigator leads to the implication that certain negative attitudes of experienced teachers toward statistics may be acting as an obstacle to their own professional growth, especially since statistics could be used as a conceptual tool in better understanding what they observe in their daily work.

One can wonder if there is a conceptual and affective gap between the teacher of measurement and the students of measurement in general. It certainly exists for statistics. The college teacher is deeply committed to his discipline, but the college student, even when he learns what he is told to learn, may not understand *why* it is important to learn it. Perhaps measurement teachers should contrive more ingenious ways to demonstrate the ultimate usefulness of certain competencies *as they are being learned*, rather than to trust to luck that they will be learned long afterwards.

In the opinion of the Principal Investigator the pre-service tests and measurements course itself could be improved in a number of ways, e.g.: (a) use of more and better audio-visual aids; (b) more laboratory and field experiences; (c) more meaningful presentation of material; (d) improved evaluation of achievement; (e) establishment of minimum or optimal standards for measurement courses. The above may wound the ego of some professors who teach measurement courses and who judge that they are doing as well as they should do. The Principal Investigator is of the opinion that teaching can always be improved. Alternatively, one may also conceive of improving the learning and emphasizing the independent role of the student in an improved self-instructional environment.

There is a strong implication that, since some measurement is needed by all teachers and since students who have taken coursework show superior competency, a measurement course should be made compulsory for every prospective teacher. Needless to say, it needs to be an interesting and meaningful compulsory course.

Perhaps "quality control," previously recommended for the training institution to insure actual development of measurement competencies, should also be utilized by State certifying agencies for the same purpose.

Several needed lines of research as a follow-up to this project have been conceived by the Principal Investigator and are suggested below.

There is a need to close the gap which exists between the teacher at the infra-college level and the professor or test specialist at the college level. Researchers from the colleges and universities should talk more with teachers and obtain job description and observational data on how teachers use measurement competencies. From this would come a refined definition of the competencies which are actually needed. There was some feedback from teachers in defining the competencies in this study. However, more is needed. Two principal avenues might be used to gather such data. First, professors in measurement courses at the universities could initiate the needed increased rapport with experienced teachers in their own classes on campus. Secondly, the researcher could go out into the field and through in-service courses, institutes, workshops, or small research projects involving discussion, interview, actual observation, etc., sample the teacher's own on-the-job behavior.

There is a need to develop better tests of measurement competency. It will not be enough to produce more items of the same type as the ones which have been used in this study, in previous studies, and in courses. There are some technical problems which need basic research. Among these is the problem concerned with making the items measure achievement status correlated with certain defined experiences and free of the influence of mental ability. Furthermore, items which measure change over a period of time need to be developed. Newer item types should be exploited in measurement of measurement competency. Among these might be situational tests, in-basket tests, more interpretive items which present pictorial or tabular background material, and oral examinations on a small scale as time allows. The nature and extent of guessing could well be studied and attempts made to assess it and compensate for it. Whereas certain topics in this study had only one, two, or three items relevant to each on the *MCT*, depth studies could be made with a subtest of a large enough number of items all of which are related to the same topic in order to insure content and construct validity, and to measure different levels of sophistication. For example, the need for low intercorrelations among subtests in a battery could be treated at a low level of simply memorizing a rule and citing it or recognizing its applicability. On a higher level it could be treated in terms of the rationale for the rule. On still a higher level, one could test for the theoretical basis, perhaps bringing in factor analysis concepts.

As an adjunct to the research activities suggested above there are some dissemination activities which come to mind.

Perhaps one avenue which would be most potent in improving the teaching of measurement would be to place in the professor's hands an instructor's handbook on improving the measurement course which would far transcend any of the current instructor's manuals which accompany specific textbooks in measurement. Such a handbook would benefit from the results of the present study and from any follow-up studies. Such a publication would not be easy to produce. It would take considerable time, expense, and effort of a large number of professional people.

Still another avenue which should be seized upon opportunistically might be to use the current trends toward increasing the quality and quantity of educational research in the field (often under the name of "evaluation" of the outcome of a funded program or project, such as Title I and Title III under P.L. 89-10) as a reason for improving the sophistication of teachers and then take steps to both influence the attitudes of teachers more favorably and to instruct them in the understandings they need in order to cooperate with more research-oriented colleagues. The increasing number of research directors in school districts or consortia among several districts should act as catalytic agents in assisting teachers along these directions.

The above suggestions about dissemination refer to work with in-service teachers and may seem beyond the scope of this project on pre-service preparation. However, improvement of measurement competency of student teachers will be relatively easier to accomplish than improvement for experienced teachers. Therefore, it was necessary to generalize to the in-service status.

Chapter V

Summary

The Problem

Evaluation of outcome of instruction is generally recognized as an important role of all teachers. However, relatively little emphasis has been devoted to developing the evaluative role in teacher training as contrasted to the emphasis upon instructional competency. There is ample evidence for the foregoing point of view. A minority of teacher-training institutions require a measurement course for their students and a minority of states require a measurement course for certification.

The Committee on Pre-Service Preparation of Teachers in Measurement of the National Council on Measurement in Education, as a result of several years of preliminary study felt the need to survey the measurement competency of beginning teachers with a view toward upgrading their preparation. This study arose from activities and convictions of members of that committee.

Methodology

The first phase of the project was to define the set of competencies which would be needed by beginning teachers. This phase consisted of developing the *Checklist of Measurement Competencies* from an existing outline of the NCME Committee. The *Checklist* was submitted to a national sample of experts (teachers, administrators, professors, and various specialists). Summary statistics from experts' response indicated judged importance of various competencies for beginning teachers.

The second phase was to construct and use an objective test, namely, the *Measurement Competency Test* (the MCT). Item selection was guided in large part by *Checklist* responses. The test was administered to samples of graduating seniors in eighty-six teacher-training institutions in the spring of 1964. The total usable sample was 2,877 students. In addition to the MCT a biographical questionnaire was administered concurrently to collect data on personal characteristics, coursework in high school and college, and the college curriculum followed.

The third phase consisted of a follow-up of the seniors two years after graduation. The *MCT* was administered to a sample of those who would cooperate concurrently with a questionnaire on intervening experiences during the two years.

Results

Quantitative results of the first phase, the definition of measurement competencies, yielded a ranking of behaviorally stated competencies so that the least important ones could be minimized or eliminated and the remainder weighted in emphasis for use in a table of specification for the *MCT*. Qualitative results showed that experts considered most *Checklist* competencies listed to be important. A few competencies were thought to be virtually non-essential for beginning teachers. Statistics competencies (especially the more abstract rather than applied) were rated low by teachers as compared with the other experts. It was felt that some competencies, although not important for beginning teachers, should be acquired by experienced teachers.

The second phase yielded data from the *MCT* and a second biographical questionnaire.

Among *Questionnaire* results were the following: Nearly all students had had at least two years of high school mathematics, while over half had had three or four years; high school science showed similar results; the college mathematics picture was different in that less than half had taken more than five semester hours; a majority had taken six to fifteen hours of college science; college psychology showed a majority taking from six to twelve hours; professional education courses far outweighed other categories, with most students having taken over twenty-one semester hours; elementary was the most popular major teaching field, accounting for almost a third; less than one-half had taken as much as one full course in tests and measurements; only one in ten had

had at least one course in statistics; about half of those who had had a course in tests and measurements had taken it at least one year previously; other variables, which will not be summarized here, were when student teaching was taken, transfer pattern, and teaching experience.

MCT mean scores for Forms A and B on the first testing were between 40 and 50 per cent of the maximum possible score. Means for the eighty-six institutions showed a very large variation, about two standard deviations. The forms did not show a close parallel.

MCT scores were tested against a number of student and institutional variables. Mean *MCT* scores for institutions were not related to type of control, type of institution, geographical region and selectivity. The *MCT* was found to be related to teaching field, amount of tests and measurements taken, amount of statistics taken, and verbal ability. It was found to be unrelated to sex.

The follow-up was carried out on 541 persons out of the original 2,877. The amount of gain for the groups on either form of the *MCT* was slightly more than two tests score point, about one-third of a standard deviation. When gain was tested against six variables only one showed significance. The five non-related variables were teaching field, amount of statistics taken, teaching experience, in-service training, and graduate study. The one related variable was amount of tests and measurements and the relation was an inverse one. The more tests and measurements taken, the smaller the gain.

Conclusions

The most important conclusions drawn were as follows:

1. There is general agreement on importance of some measurement competencies for teachers, but disagreement as to how and when teachers should acquire them.

2. There is a strong bias against statistics among some teachers.

3. Beginning teachers do not demonstrate a very high level of measurement competency as defined by project staff and experts. Completion of a course in measurement results in a modest superiority of competency as did majoring in certain teaching fields.

4. During the two years following graduation, persons from teacher training programs show a very small gain in measurement competency. Intervening experiences, such as graduate study, in-service training or teaching, did not explain any of the gain found.

5. Verbal ability was significantly related to measurement competency and teaching field.

Implications

From the above conclusions, several implications are suggested.

Further study is needed of consensus as to competencies needed for teachers of specified characteristics and in specified circumstances. Perhaps if local norms were developed for a test of measurement competency and differentiated for various teaching fields,quality control of measurement competency could be assured during training.

Two possible obstacles impeding improvement of the measurement competency level of student teachers may be (1) the lack of deep commitment to problems and practices in evaluation, and (2) negative attitude toward statistics.

Perhaps professors in measurement courses should contrive more ingenious ways to demonstrate the ultimate usefulness of certain competencies as they are being learned, rather than to trust to luck that they will be learned long afterward.

Breakthroughs are needed to improve the efficiency of pre-service training of teachers in their evaluative role. More meaningful and measurement-relevant experiences must be provided both during the pre-service and in-service periods by imaginative instructors using better teaching aids.

It may even be desirable to add evidence of measurement competency as an additional requirement for certification.

Two general lines of needed research were suggested. First, there is a need to close the gap which exists between infra-college level teacher and the professor. Secondly, there is a need to develop better tests of measurement competency.

Two general lines of dissemination activities were suggested. First, a handbook for the measurement professor transcending all extant ones could be produced if the necessary money and effort were expended. Secondly, efforts toward raising measurement competency could well parallel and could benefit current efforts to improve evaluation of funded projects in the schools.

As a final note it seems apparent that the high levels of measurement competency desirable for the teacher to play his evaluative, as well as his instructional role have not materialized from traditional training practices. If it is important enough, then the findings of this study should be implemented through efforts to improve training practices.

REFERENCES

Allen, Margaret E. "Status of Measurement Courses for Undergraduates in Teacher-Training Institutions." *13th Yearbook, National Council on Measurement in Education.* New York: the Council, 1956. Pp. 69-73.

Armstrong, W. Earl and Stinnett, T.M. *A Manual on Certification Requirements for School Personnel in the United States.* Washington, D.C.: National Education Association of the United States, 1962.

Astin, Alexander W. *Who Goes Where to College?* Chicago: Science Research Associates, 1965.

Byram, Harold M. *Some Problems in the Provision of Professional Education for College Teachers.* Teachers College, Columbia University of Contributions to Education, No. 576, New York: Bureau of Publications, 1933.

Cass, James and Birnbaum, Max. *Comparative Guide to American Colleges.* New York: Harper and Row, 1964.

Conant, James B. *The Education of American Teachers.* New York: McGraw-Hill, 1963.

Davis, Robert A. "The Teaching Problems of 1075 Public School Teachers." *Journal of Experimental Education* 9: 41-60; September 1940.

Ebel, Robert L. "Some Tests of Competence in Educational Measurement." *17th Yearbook, National Council on Measurements Used in Education.* Ames, Iowa: the Council, 1960.

Ebel, Robert Q. (Chairman); Engelhart, Max D.; Gardner, Eric F.; Gerberich, J.R.; Merwin, Jack C.; and Ward, Annie W. "Multiple-Choice Items for a Test of Teacher Competence in Educational Measurement." (Committee of the National Council on Measurement in Education.) Ames, Iowa: National Council on Measurement in Education, 1962.

Hastings, J. Thomas. *The Use of Test Results.* (U.S. Office of Education Cooperative Research Project No. 509), Urbana, Illinois: Bureau of Educational Research, University of Illinois, 1960.

Hastings, J. Thomas; Runkel, Philip J.; and Damrin, Dora. *Effects on Use of Tests by Teachers Trained in a Summer Institute.* Vol. 1. U.S. Office of Education Cooperative Research Project No. 702. Urbana: Bureau of Educational Research, University of Illinois, 1960.

Lindquist, E.F. *Design and Analysis of Experiments in Psychology and Education.* Boston: Houghton Mifflin Company, 1956.

Noll, Victor H. "Requirements in Educational Measurement for Prospective Teachers." *School and Society* 82: 88-90; Sept. 17, 1955.

Noll, Victor H. "Pre-service Preparation of Teachers in Measurement." Measurement and Research in Today's Schools: *Report of Twenty-Fifth Educational Conference Sponsored by the Educational Records Bureau and the American Council on Education.* Washington, D.C.: American Council on Education, 1961. Pp. 65-75.

Noll, Victor H. "Problems in the Pre-Service Preparation of Teachers in Measurement." *18th Yearbook, National Council on Measurement in Education.* Ames, Iowa: the Council, 1961. Pp. 35-42

Noll, Victor H. and Saupe, Joe L. *Instructor's Manual to Accompany Introduction to Educational Measurement* (1st Ed.). Boston: Houghton Mifflin Co., 1959.

Teacher Supply and Demand in Public Schools. Washington, D.C.: National Education Association, 1964.

Thorndike, Robert L. and Hagen, Elizabeth. *Teacher's Manual for Measurement and Evaluation in Psychology and Education.* (2nd Ed.). New York: John Wiley and Sons, Inc., 1955.

Appendix A

TENTATIVE OUTLINE OF NEEDED COMPETENCE IN MEASUREMENT FOR PROSPECTIVE TEACHERS

I. Standardized Tests

   A. As Contrasted to Teacher-Made Tests
      1. In construction and norming
      2. Importance of proper administration
      3. Importance of security

   B. Achievement Tests*
      1. Specific subjects and areas
      2. Survey batteries
      3. Diagnostic

   C. Intelligence and/or Aptitude Tests*
      1. Group tests
      2. Individual tests
      3. Aptitude batteries
      4. Special aptitudes

   D. Affective Test--Self Reports*
      1. Interest inventories
      2. Measures of attitudes and values
      3. Personality inventories
      4. Projective techniques

   E. Observational and Rating Techniques*
      1. Ratings
         a. Peer
         b. Supervisor
      2. Sociometric procedures
      3. Observations and anecdotal records

*For each type of measurement device listed, teachers should be aware of the following:

1. Purpose for which device is useful
2. Strengths and weaknesses of the device
3. Skills needed to use and interpret the device
4. Implications of the device for the total educational program

II. Construction and Evaluation of Classroom Tests

A. Formulate Objectives in Behavioral Terms which can be Measured

B. Devise Items to Measure Objectives

1. Knowledge of different measuring and evaluating techniques
2. Knowledge of different types of items
3. Skill in constructing test items of different types

C. Knowledge of Good Format and Arrangement of Tests, Answer Sheets, etc.

1. Arrangement of items, directions on tests, format recording or marking of answers, etc.
2. Forms, uses, advantages and disadvantages of answer sheets
3. Directions for administering tests
4. Directions for scoring tests

D. Administering a Test

1. Establishing good rapport
2. Seating, physical conditions of the room
3. Distributing materials, extra supplies, collecting materials

E. Scoring the Test

1. Arrangement of test items for scoring consumable tests
2. Types of scoring keys
3. Principles of efficent, accurate scoring

F. Evaluating the Test as a Measuring Instrument

1. Validity
2. Reliability
3. Item analysis
    a. Difficulty
    b. Discrimination

G. Sources of Information about Tests

1. Periodicals
2. Books
3. Bulletins
4. Test manuals

H. Recording and Interpreting Test Results
  1. Cumulative records
  2. Reporting and interpreting to pupils
  3. Reporting and interpreting to parents

III. Uses of Measurement and Evaluation

A. Classification
  1. Homogeneous grouping--classification with a grade
  2. Classification by grade or age

B. Diagnosis
  Identifying strengths and weaknesses in pupil's learning and in teaching

C. Counseling and Guidance
  1. Educational
  2. Vocational
  3. Personal and social

D. Marking
  Use of test results in evaluating pupil achievement

E. Identification and Study of Exceptional Children
  1. The handicapped
  2. The gifted

F. Curriculum Study and Revision
  1. Evaluation of courses and curriculums
  2. Evaluation of curriculum experimentation

G. Interpreting Schools to the Community
  1. Inter-school comparisons
  2. Comparison with national norms
  3. Interpretation of pupil marks

H. Improvement of Staff and Educational Research
  1. Help teachers in studying own methods, effectiveness
  2. Improving pupil-teacher relationships, rapport

3. Evaluation of instructional aids, programmed learning, etc.
4. Selection of staff
5. In-service education

IV. Statistical Concepts

As in all levels of learning, there are varying degrees of proficiency. This is also true insofar as statistical concepts for the beginning teacher is concerned. For this reason, we have classified the degree of proficiency or understanding required into the following:

Level of understanding and ability to compute		
1. Frequency distribution		
2. Measures of central tendency i. mean ii. median		
3. Measures of variability or scatter i. range	Measures of variability or scatter i. standard deviation ii. quartile deviation	
4. Percentiles and percentile rank	Standard scores concept	
5. Ratio I.Q. Deviation I.Q.		
6. Simple item analysis: Concept of discrimination and difficulty	Measure of relationship: Coefficient of correlation	Coeff. of correlation i. Pearson product-moment ii. Rank-order
7.	Norms	
8.	Simple bivariate expectancy table	
9.	Concept of error in measurement	Error in measure- i. std. error of mean ii. std. error of estimate iii. std. error of measurement iv. errors of technique v. errors of measurement vi. errors of sampling
10.	Concept of validity	Types of Validity
11.	Concept of reliability	Types of Reliability

Appendix B

LOYOLA UNIVERSITY

*Lewis Towers * 820 North Michigan Avenue, Chicago 11, Illinois * WHitehall 4-0800*

November 27, 1963

Your name has been given to me as one well qualified to speak in your field and as one interested in its advancement. You were recommended as one who could provide judgments as to what a teacher *should* know about tests and measurements. As you can see by the enclosed SUMMARY OF PROPOSED RESEARCH, I am directing a Cooperative Research Project to study the preservice preparation of teachers in educational measurement. We are presently implementing Objective (1) of the study, namely, "To develop a definition of competencies in educational measurement needed by teachers." Could you please help us by completing the enclosed CHECKLIST OF MEASUREMENT COMPETENCIES, so that we may be assured of an adequate cataloging of what teachers should know about measurement. It is hoped that the study may point towards ways of improving the preparation of teachers at all levels.

May we please receive your responses to the Checklist on or before December 17th. A stamped, self-addressed envelope is enclosed for your convenience. Needless to say, your replies will remain confidential. Your name is an optional part of your response, although we would like to have your title and classification.

Our budget does not permit us to offer you dollar-compensation. However, we will be happy to send you a summary of the results of the Checklist responses and a report on later results of the study.

Thank you for your cooperation.

Sincerely yours,

SAMUEL T. MAYO
Associate Professor of
Education & Director,
Cooperative Research
Project #2221

Enclosures: 2

Appendix C

CHECKLIST OF MEASUREMENT COMPETENCIES

Directions:

Please respond to the statements below in terms of knowledge, ability, and understanding which you believe the beginning teacher with a Bachelor's degree should possess.

Using an "X" mark, indicate whether you believe that each of the competencies "Is Essential," "Is Desirable," or "Is of Little Importance" to the work of the beginning teacher. If you do not understand some part of the statement check with an "X" in the last column at right entitled "Do Not Understand Statement." Also circle the part or parts of the statement which you do not understand. You may also wish to qualify your responses by writing in comments. If you wish to add any competencies which should have been included, feel free to do so on separate pages.

		Is Essential	Is Desirable	Is of Little Importance	Do Not Understand Statement
1.	Knowledge of advantages and disadvantages of standardized tests.				
2.	Ability to compare standardized with teacher-made tests and choose appropriately in a local situation.				
3.	Ability to interpret achievement test scores.				
4.	Understanding of the importance of adhering strictly to the directions and stated time limits of standardized tests.				
5.	Knowledge of sources of information about standardized tests.				
6.	Knowledge of general information about group intelligence tests.				
7.	Knowledge of general information about individual intelligence and aptitude tests.				
8.	Familiarity with need for and application of personality and interest inventories.				
9.	Familiarity with need for and application of projective techniques.				
10.	Knowledge of general uses of tests, such as motivating, emphasizing important teaching objectives in the minds of pupils, providing practice in skill, and guiding learning.				
11.	Knowledge of advantages and disadvantages of teacher-made tests.				
12.	Knowledge of the fact that test items should be constructed in terms of both content and behavior.				
13.	Ability to state measurable educational objectives.				
14.	Knowledge of the general principles of test construction (e.g., planning the test, preparing the test and evaluating the test.)				

		Is Essential	Is Desirable	Is of Little Importance	Do Not Understand Statement
15.	Knowledge of advantages and disadvantages of various types of objective test items.				
16.	Knowledge of the techniques of administering a test.				
17.	Ability to construct different types of test items.				
18.	Understanding and application of correction-for-guessing formula to an objective test.				
19.	Knowledge of the principles involved in scoring subjective and objective tests.				
20.	Knowledge of effective procedures in reporting to parents.				
21.	Knowledge of effective marking procedures.				
22.	Knowledge of advantages and disadvantages of essay questions.				
23.	Familiarity with the blueprint scheme for dealing with the content and behavior dimensions in test planning.				
24.	Ability to interpret diagnostic test results so as to evaluate pupil progress.				
25.	Ability to interpret the ratio formula relating CA, MA and IQ.				
26.	Familiarity with expected academic behavior of students classified in certain IQ ranges.				
27.	Ability to interpret a profile of sub-test results of standardized tests.				
28.	Knowledge of limitations of tests that require reading comprehension.				
29.	Understanding of the limitations of the "percentage" system of marking.				
30.	Understanding of the limitations of applying national norms to a local situation.				

		Is Essential	Is Desirable	Is of Little Importance	Do Not Understand Statement
31.	Ability to compare two classes on the basis of the means and standard deviations of a test.				
32.	Knowledge of concepts of validity, reliability and item analysis.				
33.	Ability to do a simple item analysis for a teacher-made test.				
34.	Knowledge of the limitations of ability grouping based on only one measure of ability.				
35.	Knowledge of limitations in interpreting IQ scores.				
36.	Familiarity with the nature and uses of a frequency distribution.				
37.	Familiarity with techniques of ranking a set of scores.				
38.	Ability to set up class intervals for a frequency distribution.				
39.	Understanding of the basic concept of the standard error of measurement.				
40.	Understanding of the nature and uses of the histogram and frequency polygon.				
41.	Understanding of the nature and uses of the mode, median and mean.				
42.	Ability to compute the mode, median and mean for simple sets of data.				
43.	Knowledge of advantages and disadvantages of the mode, median and mean.				
44.	Understanding of the meaning of the term "variability" and its connection with such terms as "scatter," "dispersion," "deviation," "homogeneity" and "heterogeneity."				
45.	Understanding of the nature and uses of the semi-interquartile range.				

		Is Essential	Is Desirable	Is of Little Importance	Do Not Understand Statement
46.	Understanding of the nature and uses of the standard deviation.				
47.	Ability to compute the semi-interquartile range for simple sets of data.				
48.	Knowledge of the approximate percentile ranks associated with standard scores along the horizontal baseline of the normal curve.				
49.	Knowledge of the percentage of the total number of cases included between + or - 1, 2 or 3 standard deviations from the mean in a normal distribution.				
50.	Knowledge of the fact that the normal curve is an ideal distribution, an abstract model approached but never achieved fully in practice.				
51.	Knowledge of the limitations of using the normal curve in practice as the fact that in large heterogeneous groups it "fits" most test data rather well and that it aids in the interpretation of test scores, but does not necessarily apply to small selected groups.				
52.	Ability to convert a given raw score into a z score from a mean and standard deviation of a set of scores.				
53.	Knowledge of the means and standard deviations of common standard score scales such as the z, T, stanine, deviation IQ and CEEB scales.				
54.	Knowledge of the common applications of standard scores.				
55.	Knowledge of how to convert from one type of standard score to another.				
56.	Knowledge of the fact that the mode, mean and median coincide for a symmetrical distribution.				
57.	Knowledge of the meaning of the terms used to designate certain common non-normal distributions such as "positively skewed," "negatively skewed," and "bimodal" distributions.				

		Is Essential	Is Desirable	Is of Little Importance	Do Not Understand Statement
58.	Knowledge of the fact that any normal distribution can be completely described in terms of its mean and standard deviation.				
59.	Ability to define the concept of correlation, including such terms as "positive correlation," "negative correlation," "no relationship" and "perfect relationship."				
60.	Knowledge of the significance of the numerical magnitude and the sign of the Pearson Product-Moment Correlation Coefficient.				
61.	Knowledge of the fact that correlation coefficients do not imply causality between two measures.				
62.	Knowledge of the fact that correlation coefficients alone do not indicate any kind of percentage.				
63.	Understanding of the meaning of a given correlation coefficient in terms of whether it is "high," "low" or "moderate."				
64.	Familiarity with the scatter diagram and the ability to make simple interpretations from it.				
65.	Knowledge of what size of correlation to expect between two given variables in terms of logical reasoning, e.g., in terms of a common factor.				
66.	Understanding of the fact that a raw score has no meaning alone and needs some context in which it can be interpreted.				
67.	Familiarity with the nature and uses of the common derived scores, viz., age scales, grade scales, percentile scales and standard score scales.				
68.	Understanding of certain concepts associated with scale theory, such as types of scales (nominal, ordinal, cardinal and absolute); translation of scores to a common scale; units of equal size; and common reference points (zero or the mean).				

		Is Essential	Is Desirable	Is of Little Importance	Do Not Understand Statement
69.	Ability to interpret raw scores from a given set of norms.				
70.	Understanding of the fact that interpretations of achievement from norms is affected by ability level, cultural background and curricular factors.				

Appendix D

Checklist Statements Ranked in Order of Mean Response

Legend for Column Headings

3 - Essential
2 - Desirable
1 - Of Little Importance

0 - Do Not Understand Statement
B - Left Blank
M - Mean Response

		Responses					
	Checklist Statement	3	2	1	0	B	M
4.	Understanding of the importance of adhering strictly to the directions and stated time limits of standardized tests.	164	19	0	0	2	2.89
3.	Ability to interpret achievement test scores.	163	19	0	2	1	2.89
10.	Knowledge of general uses of tests such as motivating, emphasizing important teaching objectives in the minds of the pupils, providing practice in skill, and guiding learning.	158	23	1	1	2	2.86
35.	Knowledge of limitations in interpreting IQ scores.	153	27	2	3	0	2.82
21.	Knowledge of effective marking procedures.	147	30	1	5	2	2.82
11.	Knowledge of advantages and disadvantages of teacher-made tests.	151	29	2	1	2	2.81
1.	Knowledge of advantages and disadvantages of standardized tests.	148	35	0	1	1	2.80
70.	Understanding of the fact that interpretation of achievement from norms is affected by ability level, cultural background and curricular factors.	147	32	3	1	2	2.79
66.	Understanding of the fact that a raw score has no meaning alone and needs some context in which it can be interpreted.	149	27	6	0	3	2.78
16.	Knowledge of the techniques of administering a test.	144	34	2	4	1	2.78
20.	Knowledge of effective procedures in reporting to parents.	144	39	1	0	1	2.77
14.	Knowledge of the general principles of test construction (e.g., planning the test, preparing the test and evaluating the test).	138	43	2	1	1	2.74
22.	Knowledge of advantages and disadvantages of essay questions.	130	53	0	1	1	2.71

	3	2	1	0	B	M
28. Knowledge of limitations of tests that require reading comprehension.	129	54	0	1	1	2.70
24. Ability to interpret diagnostic test results so as to evaluate pupil progress.	131	44	6	2	2	2.69
13. Ability to state measurable educational objectives.	127	48	4	4	2	2.68
29. Understanding of the limitations of the "percentage" system of marking.	124	51	6	4	0	2.65
34. Knowledge of the limitations of ability grouping based on only one measure of ability.	121	57	4	2	1	2.64
12. Knowledge of the fact that test items should be constructed in terms of both content and behavior.	113	44	8	17	3	2.63
30. Understanding of the limitations of applying national norms to a local situation.	118	57	6	4	0	2.61
15. Knowledge of the advantages and disadvantages of various types of objective test items.	118	59	6	1	1	2.61
17. Ability to construct different types of test items.	120	56	8	0	1	2.60
19. Knowledge of the principles involved in scoring subjective and objective tests.	112	62	5	5	1	2.59
6. Knowledge of general information about group intelligence tests.	110	68	4	1	2	2.58
67. Familiarity with the nature and uses of the common derived scores, viz., age scales, percentile scales, grade scales and standard score scales.	114	58	11	0	2	2.56
26. Familiarity with expected academic behavior of students classified in certain IQ ranges.	109	63	7	4	2	2.56
41. Understanding of the nature and uses of the mode, mean and median.	107	70	7	1	0	2.54
50. Knowledge of the fact that the normal curve is an ideal distribution, an abstract model approached but never achieved fully in practice.	112	56	15	1	1	2.53
27. Ability to interpret a profile of subtest results of standardized tests.	103	71	8	1	2	2.52

	3	2	1	0	B	M
2. Ability to compare standardized with teacher-made tests and choose appropriately in a local situation.	103	65	10	2	5	2.52
51. Knowledge of the limitations of using the normal curve in practice as the fact that in large heterogeneous groups it "fits" most test data rather well and that it aids in the interpretation of test scores, but does not necessarily apply to small selected groups.	111	53	19	1	1	2.50
69. Ability to interpret raw scores from a given set of norms.	97	65	14	6	3	2.47
32. Knowledge of concepts of validity, reliability and item analysis.	97	75	12	1	0	2.46
43. Knowledge of advantages and disadvantages of the mode, median and mean.	88	77	17	2	1	2.39
25. Ability to interpret the ratio formula relating CA, MA and IQ.	90	73	19	0	3	2.39
33. Ability to do a simple item analysis for a teacher-made test.	85	82	15	1	2	2.38
42. Ability to compute the mode, median and mean for simple sets of data.	87	75	22	1	0	2.35
36. Familiarity with the nature and uses of a frequency distribution.	79	90	15	1	0	2.34
61. Knowledge of the fact that correlation coefficients do not imply causality between two measures.	90	47	39	7	2	2.28
37. Familiarity with techniques of ranking a set of scores.	72	89	22	2	0	2.27
7. Knowledge of general information about individual intelligence and aptitude tests.	63	104	14	3	1	2.27
59. Ability to define the concept of correlation, including such terms as "positive correlation," "negative correlation," "no relationship" and "perfect relationship."	76	80	28	0	1	2.26
64. Familiarity with the scatter diagram and the ability to make simple interpretations from it.	69	87	23	5	1	2.25
54. Knowledge of the common applications of standard scores.	72	81	28	3	1	2.24

	3	2	1	0	B	M
5. Knowledge of sources of information about standardized tests.	61	106	16	0	2	2.24
46. Understanding of the nature and uses of the standard deviation.	71	79	32	1	2	2.21
39. Understanding of the basic concept of the standard error of measurement.	68	83	31	2	1	2.20
44. Understanding of the meaning of the term "variability" and its connection with such terms as "scatter," "dispersion," "deviation," "homogeneity" and "heterogeneity."	66	86	31	1	1	2.19
63. Understanding of the meaning of a given correlation coefficient in terms of whether it is "high," "low" or "moderate".	66	73	34	10	2	2.18
62. Knowledge of the fact that correlation coefficients alone do not indicate any kind of percentage.	69	65	40	9	2	2.16
23. Familiarity with the blueprint scheme for dealing with the content and behavior dimensions in test planning.	40	69	23	46	7	2.12
38. Ability to set up class intervals for a frequency distribution.	56	91	36	2	0	2.10
31. Ability to compare two classes on the basis of the means and standard deviations of a test.	41	103	37	3	1	2.02
48. Knowledge of the approximate percentile ranks associated with standard scores along the horizontal baseline of the normal curve.	36	104	39	4	2	1.98
49. Knowledge of the percentage of the total number of cases included between + or -1, 2 or 3 standard deviations from the mean in a normal distribution	44	88	50	2	1	1.96
56. Knowledge of the fact that the mode, mean and median coincide for a symmetrical distribution.	37	93	46	5	4	1.94
60. Knowledge of the significance of the numerical magnitude and the sign of the Pearson Product-Moment Correlation Coefficient.	42	66	53	21	3	1.93
18. Understanding and application of correction-for-guessing formula to an objective test.	28	99	55	1	2	1.85
58. Knowledge of the fact that any normal distribution can be completely described in terms of its mean and standard deviation.	34	80	62	5	4	1.84

	3	2	1	0	B	M
65. Knowledge of what size of correlation to expect between two given variables in terms of logical reasoning, e.g., in terms of a common factor.	21	92	54	15	3	1.80
40. Understanding of the nature and uses of the histogram and frequency polygon.	27	83	62	11	2	1.79
53. Knowledge of the means and standard deviations of common standard score scales, such as the z, T, stanine, deviation IQ and CEEB scales.	23	89	61	11	1	1.78
8. Familiarity with need for and application of personality and interest inventories.	21	96	62	5	1	1.77
57. Knowledge of the meaning of the terms used to designate certain common non-normal distributions such as "positively skewed," "negatively skewed," and "bimodal" distributions.	26	85	68	5	1	1.76
55. Knowledge of how to convert from one type of standard score to another.	20	90	70	3	2	1.72
45. Understanding of the nature and uses of the semi-interquartile range.	22	76	76	8	3	1.68
68. Understanding of certain concepts associated with scale theory such as types of scales (nominal, ordinal, cardinal and absolute); translation of scores to a common scale; units of equal size; and common reference points (zero or the mean).	17	86	75	5	2	1.67
52. Ability to convert a given raw score into a z score from a mean and standard deviation of a set of scores.	14	87	77	5	2	1.64
47. Ability to compute the semi-interquartile range for simple sets of data.	16	69	91	7	2	1.57
9. Familiarity with need for and application of projective techniques.	7	59	105	11	3	1.42

Number of Checklist Statements in Various Content Categories Rated "High," "Medium," or "Low," in Terms of Mean Response

Content Category	*Ratings				Identifying Item Numbers
	High	Medium	Low	Totals	
I. Standardized Tests	7	1	2	10	1-10
II. Construction & Evaluation of Classroom Tests	7	5	1	13	11-23
III. Uses of Measurement and Evaluation	4	9		13	24-36
IV. Statistical Concepts	2	18	14	34	37-70
TOTALS	20	33	17	70	

*Legend for Ratings

Rating	Range of Means
High	2.65-2.89
Medium	2.02-2.64
Low	1.42-1.98

Appendix E

MEASUREMENT COMPETENCY TEST - FORM A

In the blank, beside each item, PRINT the letter of the answer you believe to be correct.

---

1. The essential difference between standardized and unstandardized tests lies in
   A. their validity.
   B. their objectivity.
   C. the availability of norms.
   D. the discriminatory capacity of their items.

2. Advocates of "culture fair" tests of mental ability can most justifiably criticize the Stanford-Binet because of its emphasis in measuring
   A. organization of ideas.
   B. fluency of ideas.
   C. verbal abilities.
   D. innate abilities.

3. If a student wanted to find the most appropriate achievement test in arithmetic, he should consult
   A. publishers' catalogues.
   B. Buros' Mental Measurements Yearbook.
   C. Journal of Experimental Education.
   D. the most recent texts in the teaching of arithmetic.

4. If a teacher wanted to determine how well a standardized test would measure the objectives which she had been trying to teach, it would be best for her to examine
   A. the test **itself.**
   B. critical reviews of the test.
   C. the manual for the test.
   D. recent studies in which the test had been used.

5. The type of measuring device considered to require the most technical knowledge for its administration and interpretation is
   A. a group intelligence test.
   B. a self-report personality inventory.
   C. a projective test of personality.
   D. a survey achievement battery.

6. The distinction between aptitude and achievement tests is chiefly one of
   A. purpose for which used.
   B. type of ability measured.
   C. method of measurement.
   D. breadth of content.

7. Two general **types** of achievement tests have been used in secondary grades. These are (1) tests of knowledge of content common to many textbooks, and (2) tests requiring application and interpretation. What is the current status of the two types of tests?
   A. Most current tests are of type 1 and current emphasis is in the direction of type 1.
   B. Most current tests are of type 1 but current emphasis is in the direction of type 2.
   C. Most current tests are of type 2 but current emphasis is in the direction of type 1.
   D. Most current tests are of type 2 and current emphasis is in the direction of type 2.

8. High interest inventory scores relevant to a given occupation are most likely to be predictive of
   A. success in training for the occupation.
   B. actual future employment in the specific occupation.
   C. degree of success within the occupation.
   D. satisfaction with the occupation, assuming employment and requisite ability.

9. Scores on standardized intelligence tests are based on the assumption that all pupils
   A. have had some experience with such tests.
   B. have had some formal schooling.
   C. have had similar backgrounds of experience.
   D. are unfamiliar with the test material.

10. Which one of the following scores appearing in a student's record would be most meaningful without further reference to the group?
   A. 23 items correct in an English test of 40 items.
   B. 30 items wrong in an algebra test of 50 items.
   C. 100 words per minute in a typewriting test.
   D. Omitted ten items in each of the English and algebra tests.

11. The Navy reports aptitude test results in terms of standard scores with a mean of 50 and a standard deviation of 10. A recruit with mechanical comprehension score of 65 is a candidate for machinist training. On the basis of this score he would be judged
   A. a very promising candidate.
   B. slightly above average.
   C. average.
   D. slightly below average.

For each of the following paired items, PRINT A, B, C, or D in the space provided to indicate that the first item is

A greater than the second
B less than the second
C definitely equal to the second
D of uncertain size with reference to the second

12.	Usefulness of survey achievement batteries in providing data useful in guidance on the high school level.	.....	Usefulness of survey achievement batteries in providing data useful in assigning grades on the high school level.
13.	The amount of structuring in a non-projective personality test.	.....	The amount of structuring in a typical projective personality test.
14.	Usefulness of a vocational interest inventory in predicting vocational success.	.....	Usefulness of a vocational aptitude test in predicting vocational success.
15.	Importance of the physical conditions of the room upon test performance.	.....	Importance of health factors upon test performance.

In the blank, beside each item, PRINT the letter of the answer you believe to be correct.

16. It is more appropriate to discuss the mental stanine of a child with a parent than the child's I.Q. because
   A. the stanine is a more valid measure of intelligence.
   B. the I.Q. appears more precise than it actually is.
   C. mental stanines are more highly correlated with achievement.
   D. parents are better kept in doubt with reference to the child's ability.

17. What is the major argument for using unstructured essay exercises in tests given during instruction?
   A. Unstructured exercises insure that students attack the same problems.
   B. Teacher insights with reference to student thought patterns and attitudes are promoted.
   C. Course marks are more valid measures of student ability.
   D. Such exercises best stimulate students to write well-organized essay answers.

18. Why is it most desirable to use such words as "contrast," "compare" and "criticize" in formulating essay exercises?
   A. Such words are readily understood by students.
   B. Such words tend to characterize unstructured exercises.
   C. Such words stimulate students to recall relevant facts.
   D. Such words tend to characterize thought rather than fact questions.

19. How reliably can answers to essay questions be evaluated?
    A. It is impossible to evaluate them reliably enough to justify the use of this form.
    B. Under certain conditions they can be evaluated reliably, but the process is likely to be difficult and costly.
    C. They can be evaluated reliably with great ease if certain simple precautions are observed.
    D. They are ordinarily evaluated with as much reliability as are objective tests.

20. Which of the following types of items is well adapted to evaluating student knowledge of numerous technical terms?
    A. True-false.
    B. Multiple-choice.
    C. Matching.
    D. Analogy.

21. The term objective, when used to label an educational test, describes
    A. a characteristic of the scoring process.
    B. a typographic feature of the test.
    C. the degree of standardization of the test.
    D. the content limitations of the questions.

22. Sue answered correctly 25 out of 50 items on an arithmetic test. What interpretation can be made of Sue's performance on the test?
    A. Sue placed at the 50th percentile.
    B. Sue needs remedial work in arithmetic.
    C. Sue knows about one-half of the material in arithmetic taught in her grade.
    D. No interpretation of the score is possible on the basis of the information given.

23. Which of the following is a poor suggestion for the construction and use of essay examinations?
    A. Restrict the use of the essay examination to those levels of knowledge to which it is best adapted.
    B. Make definite provisions for teaching pupils how to take examinations.
    C. Increase the number of questions asked but restrict the possible answers.
    D. All of these are good suggestions.

24. Problems arise in attempting to develop measures of ultimate goals mainly because
    A. measurement methods have not given proper weight to all goals.
    B. teachers have been reluctant to depart from traditional testing methods.
    C. group norms with which to compare results are not available.
    D. such goals concern behavior not usually observable under classroom conditions.

25. Which of the following is an untrue statement about instructional goals?
    A. The worth of a goal is determined by its measurability.
    B. A two-way chart helps to relate content to educational goals.
    C. One test can usually measure only a few goals.
    D. Content and method vary directly with goals.

26. Why should behavioral objectives as contrasted with content objectives best be restricted in number?
    A. To facilitate organization of a course.
    B. To promote their operational definition.
    C. To enable a teacher to keep them constantly in mind during instruction.
    D. There are few basic factors in human ability.

27. "Washington, D.C., is the most important city in the United States." Why is this a poor true-false item?
    A. It is ambiguous.
    B. It is too easy.
    C. It is too brief.
    D. It is too factual.

28. "Philadelphia was the capital and largest city in the United States for a number of years." Why is this a poor true-false item?
    A. It is ambiguous.
    B. It involves more than one idea.
    C. It does not have a good answer.
    D. It is too long.

29. "The capital of New York State is
    1. Albany.
    2. Buffalo.
    3. Chicago.
    4. New York City."

What would be the best change to make in this item?
- A. Add the word "at" to the stem.
- B. Rewrite stem to read "Which city is the capital of New York State?"
- C. Replace "Chicago" with "Rochester."
- D. Replace "New York City" with "Syracuse."

30. "In the United States, ________ are elected for ________ and ________ for ______."

What would be the best way to revise this item?
- A. Replace the first blank by "senators" and the third blank by "representatives."
- B. Insert the word "years" after the second and fourth blanks.
- C. Insert the word "all" before the first and third blanks.
- D. Make changes A and B.

31. Validity is determined by finding the correlation between scores on
- A. the even numbered items on a test and the odd numbered items on that test.
- B. one form of a test and another form of that same test.
- C. a test and some independent criterion.
- D. two administrations of the same test.

32. What is most wrong with the statement, "This test is valid."?
- A. The statement does not specify what the test is valid for.
- B. The word "valid" is vague. A numerical coefficient should be given.
- C. A test does not show validity or lack of it.
- D. The statement is meaningless, since it does not specify the conditions of administration.

33. For determining reliability, for retesting doubtful cases, or for measuring growth, it is most useful to have
- A. equivalent forms.
- B. adequate norms.
- C. objectivity and interpretability.
- D. logical and empirical validity.

34. If the reliability of an arithmetic test is .50, and if the length is doubled, the reliability would
- A. increase.
- B. decrease.
- C. remain the same.
- D. change in some indeterminate way.

35. A spelling test is given twice within a few days to a third-grade pupil. The first time he receives a second-grade rating. His second performance puts him at the fourth-grade level. The test is probably
- A. unreliable.
- B. lacking in validity.
- C. not objective.
- D. one easily remembered.

36. Upon receiving intelligence test scores for her class a teacher is surprised to learn that a pupil she has always considered as "average" has an I.Q. of 84. Of the following, what is her most appropriate course of action?
- A. Check the pupil's cumulative record for the results of previously administered achievement and intelligence tests.
- B. Evaluate her attitude toward the pupil's performance in class to learn whether she has been grading him too leniently.
- C. Discuss the test results with the pupil to learn whether he was ill on the day of the test.
- D. Recognize that the pupil is achieving far beyond his capacity and encourage him to continue.

37. What is the chief obstacle to effective homogeneous grouping of pupils on the basis of their educational ability?
A. Resistance of children and parents to discriminations on the basis of ability.
B. Difficulty of developing suitably different teaching techniques for the various levels.
C. Increased costs of instruction as the number of groups increases and their average size decreases.
D. Wide differences in the level of development of various abilities within individual pupils.

38. A diagnostic test which provides the teacher with a profile of scores is of little value unless
A. the sub-tests which make up the profile are quite reliable.
B. the test has reliable norms.
C. the test has been shown to be a valid predictor of future achievement.
D. the scores are reported in terms of percentile ranks.

39. Peter is exactly 10 years old. His mental age is 12 years 6 months. What is his ratio I.Q.?
A. 80
B. 95
C. 125
D. None of the above.

40. In order to compute a correlation coefficient between traits A and B, it is necessary to have
A. measures of trait A on the group of persons, and of trait B on another.
B. one group of persons, some who have both A and B, some with neither, and some with one but not the other.
C. two groups of persons, one which could be classified as A or not A, the other as B or not B.
D. measures of traits A and B on each person in one group.

41. Test norms are most satisfactory when the sample of pupils or students used in establishing the norms
A. consists of nearly all pupils or students taking the test prior to the time the norms are published.
B. is representative of a clearly defined population with which it is appropriate to make comparisons.
C. ranges over all the grade levels in which the test is likely to be used.
D. includes all schools volunteering to participate in the standardization testing.

42. A good diagnostic test most differs from a good survey achievement test in
A. reliable and valid measurement of skills.
B. identifying causes of weaknesses.
C. possessing equivalent forms so that growth in achievement can be measured.
D. identifying pupils whose achievement is unsatisfactory.

43. Item difficulty values (percents of correct responses to each test item) are useful in
A. evaluating attainment of instructional objectives.
B. arranging items in order of difficulty.
C. revising a series of items.
D. accomplishing all of the above.

44. On a given test item, 30 per cent of the top fourth of the pupils marked the correct answer, and 70 per cent of the lowest fourth responded correctly. The discriminating power of the item is
A. decidedly negative.
B. slightly negative.
C. definitely positive.
D. almost perfect.

45. The State of X has a state-wide testing program. As a basis for revising the objective examination in science, a set of papers from the top and bottom quarter of the total group tested was analyzed. The per cent passing each item was determined. Other things being equal, which of the following items would one be most likely to keep in the test?
A. Top quarter -- 98%, bottom quarter -- 92%
B. Top quarter -- 80%, bottom quarter -- 40%
C. Top quarter -- 70%, bottom quarter -- 75%
D. Top quarter -- 25%, bottom quarter -- 10%

In the blank, beside each item, PRINT the letter to indicate that the item correctly refers to

A the mean
B the median
C the standard deviation
D the quartile deviation
E more than one of the above

Be sure to consider the possibility that "E" is the correct answer.

46. Is the point on the scale of measurement above which and below which there are fifty per cent of the cases.

47. An example of a measure of "central tendency."

48. Is especially useful as an average where a distribution of test scores includes a number of extremely high scores or extremely low ones.

49. Can be used in comparing their performance on a test of mental ability if computed for two different groups.

50. When computed from a frequency distribution, it is necessary at one stage to multiply by the number of units in a class interval.

51. Is represented by a distance of 10 T-score units, 2 stanine units and one z-score unit.

In the blank, beside each item, PRINT the letter of the answer you believe to be correct.

52. In the set of scores: 27, 50, 13, 5, 46, 34, 63 the median is closest to
    A. 29
    B. 34
    C. 35.4
    D. 36.5

53. Scores on standardized tests used in the elementary schools are most often converted to grade scores, for example, 4.6 or 7.3 rather than to percentile ranks. On the high school level the scores are usually converted to percentile ranks. Why?
    A. Differences in percentile ranks are in terms of equal units of ability.
    B. Grade scores assume common educational experience over the years; percentile ranks do not.
    C. Percentile ranks are necessarily more reliable than grade scores.
    D. Percentile ranks can more easily be converted to percent marks.

54. Which of the following types of derived measures is least used at the present time?
    A. Achievement quotient.
    B. Grade score.
    C. Intelligence quotient.
    D. Scaled score.

55. Find the mean of a grouped frequency distribution if the interval is 5, the arbitrary origin was taken at 25, the sum of the deviations about the arbitrary origin is 10 and the number of cases is 50.
    A. 24
    B. 25
    C. 26
    D. 27

56. A student scores 35 on a vocabulary test. The mean for the class is 37.3 and the standard deviation is 8.4, His z-score is
   - A. .27
   - B. .23
   - C. -.27
   - D. -.44

57. What does the percentile equivalent of a raw score indicate?
   - A. The per cent of a group making scores above the mid-point of that raw score interval.
   - B. The per cent of a group making scores between the upper and lower limits of that raw score interval.
   - C. The per cent of a group making scores lower than the mid-point of that raw score interval.
   - D. The per cent of items of the test which must be answered correctly to get that raw score.

58. In a particular situation the frequency distribution of scores on a standardized test is found to be approximately normal. This should be regarded as
   - A. common and highly desirable.
   - B. common but not especially desirable.
   - C. rare and highly desirable.
   - D. rare and not especially desirable.

59. If a certain test is taken by a group of high school seniors, and is found to correlate .62 with freshman grades received in college by these same seniors, one can say that
   - A. the test is a valid predictor of college aptitude.
   - B. the test is not a reliable measure of college success.
   - C. approximately two-thirds of those taking the test will be successful in college.
   - D. students who score lower than 62 will be unsuccessful in college.

60. The standard error of measurement is a numerical figure which indicates
   - A. the number of points a student's test score is in error in relation to the score he should make.
   - B. the number of points the mean score for the test is in error.
   - C. a range of scores within which the student's true score most probably falls.
   - D. the reliability of the test norms.

When you have finished the test and questionnaire, place the booklet in the enclosed self-addressed, postage-free envelope provided. Thank you for your cooperation.

Appendix F

MEASUREMENT COMPETENCY TEST - FORM B

In the blank, beside each item, PRINT the letter of the answer you believe to be correct.

1. Which of the following types of norms is least effective on the high school level?
   A. Percentile ranks.
   B. Stanines.
   C. T-scores.
   D. Grade scores.

2. The standard deviation of I.Q.'s on the Binet scale of a representative sample of white urban school children has been found to be about 16. This means that approximately 34% of the cases will have I.Q.'s between
   A. 92 and 108
   B. 84 and 116
   C. 84 and 100
   D. 100 and 132

3. A graphical device showing the distribution of scores on a single test is called a
   A. scattergram.
   B. histogram.
   C. line graph.
   D. frequency table.

4. Under a scattergram there is a notation that the coefficient of correlation is .06. This means that
   A. most of the cases are plotted within a range of 6% above or below a sloping line in the diagram.
   B. plus and minus 6% from the means includes about 68% of the cases.
   C. there is a negligible correlation between the two variables.
   D. most of the data plotted fall into a narrow band 6% wide.

5. A teacher is in the habit of giving his geometry students a weekly test. In the middle of the school year, six of the students in his class transfer to another school. For the remaining students, which of the following will probably show the greatest amount of change?
   A. The raw score they make on the weekly tests.
   B. Their rank in class as determined by the weekly tests.
   C. The average weekly test scores.
   D. The range of their weekly test scores.

6. In a frequency distribution representing a group of 50 individuals, the median is in the score interval whose indicated limits are 48-52. The number of cases up to the lower limit of this interval is 18, and there are ten cases in this interval. What proportion of the 48-52 interval falls below the median?
   A. 30%
   B. 50%
   C. 70%
   D. Indeterminate from the data given.

7. A student's raw score is exactly in the middle of the range of raw scores assigned a stanine of 7. If his raw score were assigned a T-score, it would be numerically equal to
   A. 30
   B. 40
   C. 60
   D. 75

8. In a frequency distribution of 250 scores, the mean is reported as 78 and the median as 65. One would expect this distribution to be
   A. positively skewed.
   B. negatively skewed.
   C. symmetrical.
   D. normal.

9. Which of the following shows the highest degree of correlation?
   A. +.40
   B. -.20
   C. -.50
   D. -.65

10. Below are the percentile scores of four students on a standardized reading test:

Mary:	45	Tom:	90
Jane:	50	Jim:	95

What can be said about the difference in these students' achievement?
   A. The relative differences in achievement between Mary and Jane is equal to that between Tom and Jim.
   B. Tom's achievement is twice as great as Mary's.
   C. The teacher can be more certain about Jim being better than Tom than she can about Jane being better than Mary.
   D. The teacher should recognize that if the test were administered a second time, it is quite probable that Tom would do better than Jim.

In the blank, beside each item, PRINT the letter to indicate that the item correctly refers to

A the mean
B the median
C the standard deviation
D the quartile deviation
E more than one of the above

Be sure to consider the possibility that "E" is the correct answer.

11. Includes approximately 68 per cent of the cases when measured above and below the mean in a normal distribution.

12. May be obtained by summing the scores and dividing by the total number of scores.

13. Is most often confused with the "mid-score."

14. A point that is affected markedly by extremely high or low scores.

15. Is represented by a T-score of 50, a stanine of 5 and a z-score of 0.

In the blank, beside each item, PRINT the letter of the answer you believe to be correct.

16. At the end of the semester a history teacher gave his pupils an essay test on the material covered during the preceding weeks. When he graded the papers he deducted points from the total score for spelling, grammar and English usage. In so doing, he
   A. increased the accuracy of his final grades.
   B. increased the objectivity of measurement.
   C. lowered the reliability of the test.
   D. lowered the validity of the test.

17. A teacher has given four 100-item achievement tests with the following results. Which test apparently was most suitable for the group?
   A. Test I: mean, 40; range, 17-80
   B. Test II: mean, 54; range, 18-82
   C. Test III: mean, 68; range, 36-99
   D. Test IV: mean, 88; range, 62-98

18. John scored at the 60th percentile on an academic aptitude test and scored at the 57th percentile on a test of reading ability. The above data indicate that John's teacher should
   A. ignore this difference altogether.
   B. provide him with individual help in reading.
   C. motivate him to read more extensively outside of school.
   D. have him retested in reading ability.

19. The same test is given on successive days to the same class. The correlation between the two sets of scores is .95. Which conclusion concerning the scores is most defensible?
   A. They are highly reliable.
   B. They are highly valid.
   C. They are quite unstable.
   D. They are not differentiating.

20. An achievement test item is characterized by the following item analysis data where B is the keyed answer:

	A	B	C	D	E
High Group	8	47	19	15	11
Low Group	16	19	24	26	15

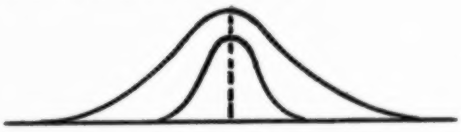
One can infer from the data given above, that this item
   A. is a relatively easy one.
   B. has distractors all needing revision.
   C. is of satisfactory discriminating power.
   D. has not been keyed correctly.

21. In tallying a frequency distribution of test scores, class intervals of 15-19, 20-24, 25-29, etc., are used. Where 22, rather than 22.5, is taken as the mid-point of the interval, the crucial assumption is that
   A. the score of 22 means a range of 22.000 to 22.999... .
   B. the score 22 means a range from 21.000... to 22.000... .
   C. the interval 20-24 means a range from 20.000... to 24.999... .
   D. the interval 20-24 means a range from 19.500... to 24.499... .

22. Quite often test manuals give analyses of the sources from which the items in a test have been drawn and include information with respect to the proportions of items relevant to different categories. This information is most useful in evaluating a test with respect to its
   A. predictive validity.
   B. content validity.
   C. construct validity.
   D. concurrent validity.

23. A deviation I.Q. indicates
   A. deviation of MA from CA.
   B. deviation of two sets of scores from the mean.
   C. the distance in standard score units of a score from the mean.
   D. relative achievement of a person in terms of standard score units.

24. The distributions shown differ in
   A. skewness only.
   B. variability only.
   C. central tendency only.
   D. both variability and central tendency.



25. In general, increasing the length of a test will make it more
   A. valid.
   B. reliable.
   C. objective.
   D. diagnostic.

26. A teacher is examining the manual for a new diagnostic reading test. In the section labeled, "Description of Test" she finds the statement: "This test provides measures of four completely independent reading skills." In the section labeled, "Test Statistics" she finds the following data on the reliability and intercorrelation of the four scores:

Reading Skills	Par. Mean.	Sent. Mean.	Vocab.	R. Speed
Paragraph Meaning	.88*			
Sentence Meaning	.80	.82*		
Reading Vocabulary	.82	.76	.88*	
Reading Speed	.78	.72	.76	.94*

*The entries in the diagonal are reliability coefficients.

On the basis of the material in the test manual, what criticism should the teacher make?
   A. The test does not measure independent reading skills.
   B. The test is highly speeded.
   C. The test is not sufficiently reliable to make comparisons between individual pupils.
   D. The correlations among the scores indicate that the test possesses little validity.

27. Because no standardized test possesses perfect reliability it is essential that the teacher regard the score which a student obtains as
   A. having little meaning unless it is very high or very low.
   B. indicating a point in the range near which the student's true score probably falls.
   C. indicating only that the student has either more or less ability than the average individual in the norming group.
   D. providing information about the student which can be used only by a thoroughly trained guidance counselor.

28. In which of the following instances is a teacher most justified in requiring all students to make test scores of 75% or better?
   A. The class is composed of above average students.
   B. The questions are essay rather than objective.
   C. The questions measure knowledge of essentials.
   D. The pupils have ample time to prepare for the test.

29. John tells his mother that he made a score of 68 on his science test. Which type of information would best help his mother to understand the meaning of his score in terms of his achievement in science?
   A. The test consisted of 90 questions.
   B. Half of the class failed the test.
   C. The mean score for the class was 65.
   D. The highest score in the class was 83.

30. Year after year the mean achievement test scores for the students in school X consistently are one year or more above the national norms. What is the most probable cause of this finding?
   A. School X is located in an upper-middle-class community.
   B. School X is staffed with expert teachers.
   C. School X is using tests that have unreliable norms.
   D. School X stresses the traditional, rather than the activity, curriculum.

31. Which of the following is a poor principle to use in marking or assigning grades?
   A. Letter grades have definite advantages over percentage grades.
   B. Marks should be based as much as possible on objective measures.
   C. Marks should indicate achievement of general as opposed to specific objectives.
   D. Status and improvement should be graded separately.

32. Objective test exercises are most likely to measure the ability of the pupils to reason if the exercises
   A. are of the recall rather than of the recognition type.
   B. are similar in form to intelligence test exercises.
   C. are of the multiple-answer rather than the true-false type.
   D. require application of facts to a novel situation or problem.

33. The use of the normal curve as a basis for assigning school marks is most legitimate when
A. a standardized test is used.
B. all of the pupils have approximately the same I.Q.
C. the marks are to be assigned to a large and representative group of pupils.
D. the average pupil scores 85 on the test used.

34. The most important advantage of the objective test over the essay test is that it
A. saves time for the teacher.
B. has higher content validity.
C. measures a greater range of instructional objectives.
D. provides for a more complete sampling of content.

35. A two-way chart is used in identifying for each item of an achievement test the topics and the behavioral objectives to which each item is relevant. The process is one of estimating the test's
A. concurrent validity.
B. predictive validity.
C. content validity.
D. construct validity.

36. In the scoring of essay examinations, all the following are generally considered desirable practices except to
A. reduce the mark for poor spelling or penmanship.
B. prepare a scoring key and standards in advance.
C. remove or cover pupils' names from the papers.
D. score one question on all papers before going to the next.

37. When is it generally desirable for the teacher to decide upon the specific format of items to be developed for a test?
A. When the evaluation plan is being developed.
B. As the very first step.
C. After the total number of questions has been decided upon.
D. After study of the specific behaviors listed in the test plan.

38. One of the best ways for a teacher to begin a study designed to formulate goals for his teaching is to
A. read the authors' prefaces of the textbooks he uses.
B. prepare an outline of the materials covered in his textbooks.
C. examine objectives formulated by other teachers.
D. discuss the problem with more experienced teachers.

39. The type of instructional outcome most difficult to evaluate objectively is
A. a concept.
B. an appreciation.
C. an attitude.
D. an understanding.

40. "Columbus discovered America in __________."
The best change to make in revising this item would be to rewrite it so as to read
A. "America was discovered by Columbus in ___________."
B. "Columbus discovered __________ in ______________."
C. "Columbus discovered America in the year of _______."
D. "__________ was discovered by Columbus in ________."

41. In which way are teacher-made tests superior to standardized tests?
A. They are more reliable for evaluating differences among very poor and very good students.
B. They provide more valid measures of the teacher's specific objectives.
C. They provide a better measure of the student's grasp of important facts and principles.
D. They are simpler to administer and score.

42. This exercise
A. is faulty because the answers are not of parallel construction.
B. is faulty because the answers do not all complete the item stem.
C. is faulty because of ambiguous phraseology.
D. is faulty because the problem is not in the item stem.

43. Measurement specialists would generally consider the practice of allowing a choice in the questions to be answered on an essay examination
A. desirable, because it gives each student a fairer chance.
B. desirable, because it permits a wider sampling of the topics covered.
C. undesirable, because it reduces the comparability of the test from student to student.
D. undesirable, because students waste too much time deciding which question to answer.

44. A science teacher is preparing a test to be used to determine knowledge of specifics from a unit of study. He should use objective rather than essay questions because they
A. avoid ambiguity, the most common fault of test questions.
B. provide a wider sampling of material.
C. are not affected by the judgment of the tester.
D. are best suited to his purpose.

45. One of the merits of arranging test items in an order of difficulty is that
A. it insures an accurate measure of consistency.
B. it encourages the pupil taking the test to continue.
C. item validity is to some extent dependent on difficulty.
D. this procedure contributes to the test's reliability.

For each of the following paired items, PRINT A, B, C, or D in the space provided to indicate that the first item is

A greater than the second
B less than the second
C definitely equal to the second
D of uncertain size with reference to the second

46. The level of ability represented by an I.Q. of 116 on the Stanford-Binet.	.....	The level of ability represented by a stanine score of 6 on the Stanford-Binet.
47. The level of achievement in reading represented by a grade score of 8.5 on the California Reading Test	.....	The level of achievement represented by a grade score of 8.5 on the Metropolitan Reading Test.
48. The justification of calling a test standardized that has been normed on 2,000 students.	.....	The justification of calling a test standardized that has been normed on 5,000 students.
49. The desirability of using standardized achievement test results for grading purposes.	.....	The desirability of using standardized achievement test results for grouping purposes.
50. Extent to which correlation of parts is justified in a test designed to measure "general" intelligence.	.....	Extent to which correlation of parts is justified in a test designed to measure several aptitudes.

In the blank, beside each item, PRINT the letter of the answer you believe to be correct.

51. In determining the grade placement of pupils new to a school, the most useful data may be obtained by administering
A. achievement tests in reading, arithmetic and science.
B. achievement tests in reading and arithmetic.
C. achievement tests in reading and arithmetic plus an attitude inventory.
D. a survey achievement battery.

52. What is usually the last step in the production of a standardized achievement test?
   A. Final revision of test items and directions.
   B. Administration to a large and representative sample of pupils.
   C. Careful evaluation of test materials by experts.
   D. Statistical analysis of test items.

53. If you were asked to serve on a committee for the purpose of selecting a standardized achievement battery for your school, or school district, you would consider each of the following but give greatest weight to
   A. unit cost per pupil tested.
   B. availability of equivalent forms.
   C. relevance to local instructional objectives.
   D. ease of administration and scoring.

54. In a battery measuring various aptitudes the subtests should have
   A. low correlations with each other and high reliability coefficients.
   B. high correlations with grade-point averages in college.
   C. negative correlations with each other.
   D. validity coefficients higher than their reliability coefficients.

55. In giving a standardized test a teacher allows too much time. This is most likely to adversely affect
   A. the reliability of the test.
   B. the validity of the test.
   C. interpretation in terms of norms.
   D. the ranking of pupils.

56. Test techniques are generally preferred to observational techniques, when both are available for the testing purpose, because the former are
   A. more apt to yield measures.
   B. perceived as a test by the student, thus more apt to be based on a motivated performance.
   C. applicable to a wider variety of personal traits.
   D. more apt to yield reliable scores.

57. If, in administering a standardized test, one departs from the exact instructions, this will probably affect most seriously the
   A. reliability of measurement.
   B. objectivity of scoring.
   C. applicability of norms.
   D. comparability of individual scores.

58. Teachers should motivate students to make the best scores they possibly can on all of the following except
   A. aptitude measures.
   B. diagnostic measures.
   C. personality measures.
   D. readiness measures.

59. If a teacher wishes to obtain a critical review of a standardized test she plans to use with her classes, she should consult the
   A. test Manual issued by the publisher.
   B. Encyclopedia of Educational Research.
   C. Review of Educational Research.
   D. Mental Measurements Yearbook.

60. In contrast to a test which is "well standardized" a poorly standardized test is one which
   A. has norms that are based on fewer than 1,000 cases.
   B. uses a norm sample that is not representative of the group for which the test is designed.
   C. consists of test questions that have not been validated.
   D. includes test questions that do not measure what they are intended to measure.

When you have finished the test and questionnaire, place the booklet in the enclosed self-addressed, postage-free envelope provided. Thank you for your cooperation.

Appendix G

QUESTIONNAIRE FOR SENIORS IN TEACHER-PREPARATION PROGRAMS

Department of Education
Loyola University, Chicago

Directions: Your responses will be a combination of written-in information and checked options. Where you are asked to "Check One," indicate your response by making an "X" in the appropriate blank. Where college coursework is called for, include concurrent courses.

1. Institution ____
(College or university where you are taking your teacher-preparation)

2. Name ____
Last Name First Name Middle Name

3. Permanent Mailing Address (where you can always be reached)

____

____

4. Age last birthday ________

5. Sex (Check one)
____1. Male
____2. Female

6. Mathematics coursework in high school (number of years) _____

7. Science coursework in high school (number of years) _____

8. (a) Mathematics coursework in college (number of credit hours) _____

(b) Type of credit hour (Check one)
____1. Quarter hour
____2. Semester hour

9. (a) Science coursework in college (number of credit hours) _____

(b) Type of credit hour (Check one)
____1. Quarter hour
____2. Semester hour

10. (a) Psychology coursework in college (number of credit hours) _____

(b) Type of credit hour (Check one)
____1. Quarter hour
____2. Semester hour

11. (a) Professional education coursework (i.e., carrying credit in a department or a school of education)--(number of credit hours) _____

(b) Type of credit hour (Check one)
____1. Quarter hour
____2. Semester hour

12. Level of your teacher preparation (Check one or two)
____1. Nursery School & Kindergarten
____2. Grades 1 - 3
____3. Grades 4 - 6
____4. Grades 7 - 8
____5. Grades 9 - 12
____6. Other (Specify) ______________________________

13. Teaching field (Check one or two and circle your major field if you check two fields).
____1. General Elementary
____2. English
____3. Mathematics
____4. Science
____5. Social Science
____6. Art
____7. Music
____8. Foreign Languages
____9. Business & Commercial
___10. Industrial Arts (Vocational)
__11. Industrial Arts (Non-Vocational)
____12. Agriculture
____13. Home Economics
____14. Physical Education
____15. Exceptional Children
____16. Speech Correction
____17. Health Education
____18. Recreation
____19. Other (Specify) __________
____

14. Where was majority of your college work completed? (Check one)
____1. At present institution
____2. At (an)other institution(s)

15. If you transferred, indicate when you transferred to present institution. (Check one)
____1. Freshman
____2. Sophomore
____3. Junior
____4. Senior
____5. Did not transfer

16. Number of years of teaching experience other than student teaching ____

17. Statistics coursework in college (Check one or a combination of "2" & "3")
____1. None
____2. Part of another course (Specify name of course(s). ________
____
____3. One full course
____4. More than one course

18. How much coursework have you had in tests and measurements?
____1. None
____2. Part of another course (Specify name of course(s). ________
____
____3. One full course
____4. More than one course

19. If you have had coursework in tests and measurements, when was it or is it being completed?
____1. Currently
____2. Last term
____3. One year ago
____4. Two years ago
____5. More than two years ago

20. Is your student teaching already completed or is it currently being taken?
____1. Already completed
____2. Currently being taken
____3. Has not been taken

Appendix H

LOYOLA UNIVERSITY

*Lewis Towers * 820 North Michigan Avenue, Chicago 11, Illinois * WHitehall 4-0800*

As part of the United States Office of Education Cooperative Research Project as described on the enclosed *Summary of Proposed Research* your institution has been selected by random sampling as a source for a sample of seniors who have had teacher preparation. We wish to test a proportion of the seniors in each of more than 100 institutions in a nationally representative sample. We know that you have a busy schedule in your institution and that time is at a premium. Nevertheless, we do feel that this project has extremely important implications for the improvement of education and particularly, for the improvement of measurement competencies of teachers and prospective teachers. We certainly hope that you will share our interest and consent to cooperate in this undertaking.

The procedures of a participating institution will be as follows: (a) Based on the number of last term seniors per institution (as yielded by the enclosed questionnaire) a proportion of seniors per institution, probably around 30 per cent, will be determined by the Project Director. (b) A roster of seniors' names or of class sections will be numbered in any arbitrary order by the institution. (c) Names of seniors finally chosen will be determined by a random sample of senior numbers furnished by the Project Director. Seniors can be tested either in regular class periods or outside the class periods on a group or individual basis. The test will be of the untimed, or power, type. It is planned that the test can be administered in approximately one hour.

Although we can offer no dollar-compensation for your trouble, we will be happy to send you a report of the test results which will be anonymous except for identifying the results of your institution to you only.

Would you please indicate on the enclosed questionnaire whether you will be able to participate in testing a sample of your seniors in April or May of 1964. Your cooperation will be deeply appreciated and will make the project more successful. If possible, would you let us hear from you in approximately a week to ten days.

SAMUEL T. MAYO, Ph.D.
Director, Cooperative Research
Project #2221

STM:bb
Enclosures: 3

Appendix I

COOPERATIVE RESEARCH PROJECT #2221

Summary of Proposed Research

Title. Pre-Service Preparation of Teachers in Educational Measurement.

Principal Investigator. Samuel T. Mayo, Ph.D., Loyola University, Chicago.

Objectives. (1) To develop a definition of competencies in educational measurement needed by teachers; (2) To develop a measuring instrument of the desired competencies; (3) To relate actual competencies of prospective teachers at time of graduation to undergraduate programs and background; (4) To relate changes in competencies during a two-year period after graduation to intervening professional experiences; and (5) To interpret findings in relation to current programs for preparation of teachers, with implications for modification.

Procedure. In cooperation with the Committee on Pre-Service Preparation of Teachers in Measurement of the National Council on Measurement on Education, a checklist based upon their *Outline of Needed Competencies* will be prepared. The definition of needed competencies will be refined from checklist data from a selected sample of measurement experts and educators. The objective test will be administered to a representative sample of graduating seniors in teacher education programs. Test data will be analyzed in terms of discrepancies between what competencies prospective teachers actually possess and those defined as needed. Test data will also be related to undergraduate coursework and background variables. A follow-up of seniors with a second testing two years after graduation will indicate changes in competencies. Such changes will be related to intervening professional experiences.

Appendix J

COOPERATIVE RESEARCH PROJECT #2221

Questionnaire for Sample of Institutions Chosen for Graduating Seniors Sample

Name of Institution ______________________________

1. Our institution will ________ / will not ____ be able to administer a test of measurement competency to a proportion of our graduating seniors in April or May of 1964.

(NOTE: The following questions are to be answered by those institutions who responded "will" to Question No. 1.)

2. Under which system does your school operate?

Quarter __________

Semester _________

Trimester ________

Other ____________

3. Would you be able to administer the test between April 15 and May 15?

Yes ______ No ______

4. What is the estimated number of last term seniors in teacher-education for the term in which testing will be done? _____________

5. What is the name and position of the person in your institution who will coordinate the local testing?

NAME ______________________________

POSITION ______________________________

ADDRESS ______________________________

Appendix K

MEMORANDUM

COOPERATIVE RESEARCH PROJECT #2221
LOYOLA UNIVERSITY, CHICAGO

To: Coordinators of testing for sample of graduating seniors in teacher-preparation

From: Samuel T. Mayo, Director of Project

Subject: Further instructions on procedures

1. First, let me extend warm thanks for your fine cooperation in our research.

2. Some of you who received our earliest version of the covering letter for the questionnaire and who were asked to administer a two-hour test will be pleased to know that the length of the test has been reduced to less than one hour.

3. Our present schedule calls for us to have the test materials in your hands sometime during the week of April 20 through April 25.

4. The original plan to draw a *strictly random* sample of a graduating class has been impractical at some institutions. Accordingly, we have had to modify procedures at such institutions. It is necessary, at this time, to ask if you can carry out the original random procedures or whether you must resort to an alternative plan. Would you please indicate on the enclosed questionnaire which sampling plan you can best carry out, and return the form to me as soon as possible. If you can sample randomly, I will send you a list of random numbers to be referred to your arbitrarily numbered list of your students. If I do not hear from you before mailout of test materials on or about April 20, I shall still enclose the list of random numbers *hopefully*.

5. So that we will know how many test booklets and answer sheets to send, we would like to know if there has been any change in the original estimate of

the number of seniors which you filled in on the questionnaire you returned. We plan to ship a quantity of tests and answer sheets equal to 40 per cent of the estimated number of seniors which you indicated. On the enclosed questionnaire, please indicate if it will be convenient for you to test this number of seniors and if our figures agree. For some of the smaller institutions, say with twenty-five or less graduating seniors, we plan to ask for a 100 per cent sample, if feasible.

6. The answer sheets which we ship will be of the IBM type. We will not ship the special electrographic pencils. However, we would appreciate your having students use the special pencils if they are available locally. If they are not, please have them use a soft pencil (preferably no harder than a No. 2), and we will go over their marks with an electrographic pencil after the answer sheets are returned.

7. If there is any other situation which we should know about which has not been caught on the questionnaire, please feel free to write in your comments at the bottom of the page, in the margins, and on the back.

Appendix L

Questionnaire for Coordinators of Senior Testing

Name and Location of Institution ____________________

____

Directions: Check one choice in each question which applies to you and also fill in the appropriate blanks.

1. Can you carry out the original plan to draw a random sample of 40 per cent of your list of graduating seniors from a set of random numbers to be furnished you?
   ______Yes
   ______No

   If you answered "no" to question 1, please answer questions 2 through 4 below.

2. Which of the following problems, if any, would you encounter in obtaining a sample representative of your graduating seniors?
   ______Not all seniors are available on campus
   ______Seniors are broken up into smaller groups according to teaching level and field
   ______Other problem (Specify) ____________________

   ____

3. Which of the following alternative sampling plans is feasible for you?
   ______Test only the seniors on campus or nearby
   ______Test about 40 per cent of a number of intact groups
   ______Other plan (Specify) ______________________

4. Please describe the characteristics of the non-random sample you plan to use in regard to any biases in relation to the total group of seniors. (e.g., Are there any biases in the elementary vs. secondary level ratio or in abilities, or teaching fields?)

5. Will the sample size you chose in question 3 above be different from 40 per cent of the total number of seniors? (According to our records you will have an estimated ____ seniors.)
______Yes (Specify) ______________________________
______No

6. What is the present number of seniors you will need test materials for, based upon either 40 per cent of the total, a sample of available seniors on campus, or 100 per cent of seniors for smaller institutions? (Check one).
______40 per cent of the written-in figure in question 5.
______40 per cent of a different figure from the one in question 5 (Specify) ______________
______A sample of present seniors on campus, the number of which is ______
______A sample of seniors from off-campus centers, the number of which is ______

Appendix M

COOPERATIVE RESEARCH PROJECT #2221

Department of Education
Loyola University, Chicago

MEMORANDUM TO TESTING COORDINATORS

1. Test Package. The package of test materials sent to you contains test booklets, answer sheets, student questionnaires, DIRECTIONS FOR TEST ADMINISTRATION, one or more stamped, addressed return envelopes, and a return postal card. You are advised to examine *all* of these materials carefully prior to the administration of the test.

2. Test Booklets. The number of test booklets included in the package is equal to either the (a) total number of your graduating seniors if yours is a very small institution, or if you requested that we test all of your seniors; (b) 40 per cent of the total number of seniors which you indicated in our questionnaire; or (c) some other number which you indicated or which we mutually agreed upon. Students are *not* to write in the test booklets. Separate answer sheets are provided for recording answers and scratch paper is permitted for calculations.

3. Answer Sheets. The answer sheets enclosed are standard IBM answer sheets with space for 150 5-option multiple-choice items. We are using only the first four options ("A through D") on most of the items, and students should avoid marking the "E" responses except in one key-list exercise in which "E" is called for. Students should carefully and legibly print in the information called for in the margin of the answer sheet as specified in the DIRECTIONS FOR TEST ADMINISTRATION. Be sure that students mark the appropriate form on the answer sheet. All students in your institution will have the same form. After the test is completed, separate the test booklets and answer sheets in the return package.

4. Student Questionnaires. Each student should complete a copy of the questionnaire. It should be possible to administer the questionnaire and test within one hour to everyone. It may also be possible

to do this in a fifty minute period. The questionnaire in tryout form was completed by almost everyone in three or four minutes. If necessary, the questionnaire could be given a different time from the test.

5. Determining of Sample of Students. One of the following procedures will apply to your particular sampling situation:

(a) If you have a relatively small graduating class, you will test 100 per cent of your group. We have drawn the line of smallness at thirty students or less.
(b) If you agreed to identify a 40 per cent *random* sample from an arbitrary listing of your students, you may determine which particular students on the list are to be tested by the use of the enclosed CHART FOR DRAWING A RANDOM SAMPLE FOR VARYING SIZES OF GRADUATING CLASS.
(c) If you indicated or if we agreed upon some other sampling procedure, you should disregard the CHART and follow the alternative procedure.

6. Report. Please report any unusual incident or actions which might affect the validity of the testing. Also indicate any difficulties encountered.

7. Precautions. It is important that:

(a) There be no loss of tests, answer sheets or questionnaires.
(b) The answer sheets and questionnaires be properly identified and marked.
(c) You be as helpful to the students in the mechanics of the test as possible without giving them any help in the actual questions.
(d) There be constant supervision of the students while tests are in progress.

8. Return of Materials. One or more manila envelopes is enclosed for the return of materials. Postage is included. They are to be sent as "Educational Material." In the case where two or more envelopes

are included, divide the weight of materials equally among the several packages.

Appendix N

Department of Education
Loyola University, Chicago

DIRECTIONS FOR TEST ADMINISTRATION

1. Announce to students that this test is part of a federally sponsored research project to determine what prospective teachers actually know about tests and measurement at the time of graduation. It is hoped that from the project may come improvement in the preparation of teachers in measurement.

2. If possible, have students complete the questionnaire first, then take the test.

3. Scratch paper is permitted, one sheet to a student. For security reasons, it will be desirable to have all sheets of scratch paper returned with the test booklets and answer sheets. The sheets of scratch paper may then be destroyed. If it is feasible, scratch paper of uniform size, color and type should be furnished by the test administrator.

4. Distribute the student questionnaires, test booklets, answer sheets and scratch paper, keeping careful account of all test materials.

5. Have students print in the following information in the spaces provided in the margin of the answer sheet:

Name (printed)
Date (in the form exemplified by "5/13/64")
School (institution)
City
Name of Test (have them print "Meas.Comp.Test.")
Part (have them print either "A" or "B" to correspond with the form on cover of test.)

6. Ask students to read the instructions on the front cover of the test booklet. Ask if there are any questions. Announce that all of the multiple-choice and key-list items will have only four options, "A, B, C and D," except for one key-list exercise which has five options.

7. In marking the answer sheet of the test, students should use an IBM pencil, if available, or a soft pencil (no. 2), otherwise. Wax pencils, colored pencils or ink pens should not be used.

8. Students should be given sufficient time to attempt all items, since it is designed as a "power test." It is estimated that a fifty-minute period should be enough time for 95 per cent or more of a group to complete both the questionnaire and the test. If it is feasible to allow more time for the slower students, this would be appreciated. If it can be done, a log of the time required for the fastest and slowest numbers of the group on the test would be appreciated.

9. If unusual incidents occur during the administration of the test, please describe them.

Appendix 0

CHART FOR DRAWING A RANDOM SAMPLE FOR VARYING SIZES OF GRADUATING CLASS

Directions: Locate the number closest to the size of your graduating class. The numbers which come before this number will indicate the students on your list who are to be tested.

For example, if your class size is 50, find 40 per cent of 50, which is 20. Locate the number closest to your class size (in this case exactly 50), which falls in the first column, and you will find that there are 20 numbers which come before 50. These 20 numbers will constitute your random sample.

1
3
5
7
11
12
16
18
20
23
26
27
31
33
35
37
41
42
46
47
50
52
55
57
63
64
66
68
70
74
76
79
82
84
85
88
90
91
97
99

100
104
106
109
111
114
115
116
122
123
127
129
131
133
136
138
142
144
148
149
150
154
156
157
162
164
166
168
171
173
175
178
182
184
187
188
192
194
196
197

200
202
205
207
211
213
216
217
220
222
226
227
231
232
235
239
242
244
246
249
252
253
255
256
261
262
267
269
271
273
277
279
283
284
287
289
290
292
297
298

300
304
306
308
312
313
315
317
320
321
326
327
332
333
335
339
340
344
345
346
352
354
357
359
360
362
365
366
370
374
376
378
380
383
386
387
390
394
396
398

402
404
407
408
410
411
415
418
423
424
427
429
432
434
435
438
440
444
448
449
452
453
455
457
460
462
466
468
473
474
475
477
483
484
486
488
493
494
497
499

Appendix P

PLEASE CHECK THE ITEMS BELOW WHICH APPLY TO YOU:

☐ The address label below is correct.

☐ My address has changed. Corrections are written in on the label below.

☐ I have entered the teaching profession.

☐ I did not enter the teaching profession. My present employment is:

☐ I would like to have a final report on the results of the study.

Remarks:

THIS IS HOW WE ARE ADDRESSING YOU:

ARE
WE
ADDRESSING
YOU
CORRECTLY?
SENATOR?
MR.?
COLONEL?
Mr. Secretary?
INSPECTOR?
SAHIB?
DR.?
EFFENDI?
YOUR HIGHNESS?
BARON?
COUNT?
YOUR LORDSHIP?
SIR?
BWANA?
YOUR WORSHIP?
MR. PRESIDENT?
CHIEF?
CHARLEY?
PRINCE?

DR. SAMUEL T. MAYO, Director
U.S.O.E. Coop. Resch. Proj. #2221
Loyola University
820 N Michigan Ave.
Chicago, Illinois 60611

RETURN REQUESTED

FIRST CLASS MAIL

ARE WE ADDRESSING YOU
CORRECTLY?

Shortly before your graduation in 1964 you completed a questionnaire and a test as part of our national research project to improve teacher preparation.

For further study we need to keep our address file current. Would you please mark the appropriate items and return the attached card as soon as possible.

A final report of results of the study will be sent to you if you wish.

Samuel T. Mayo
Director of Project

Postage
Will be Paid
by
Addressee

No
Postage Stamp
Necessary
If Mailed in the
United States

BUSINESS REPLY CARD
FIRST CLASS PERMIT No. 13444 CHICAGO, ILLINOIS

DR. SAMUEL T. MAYO, Director
U.S.O.E. Coop. Resch. Proj. #2221
Loyola University
820 N. Michigan Ave.
Chicago, Illinois 60611

Appendix Q

LOYOLA UNIVERSITY

*Lewis Towers * 820 North Michigan Avenue, Chicago 11, Illinois * WHitehall 4-0800*

April, 1966

Dear Colleague:

During the past three years, Loyola University has been involved in a research project concerned with the preparation of teachers. We have been gathering data enabling us to assess the role of courses in tests and measurements as shown in the enclosed *Summary of Proposed Research.* Specifically, we have asked the question, "How can teachers be helped in fulfilling their evaluative role?" ... Ours is the first large-scale study in evaluation skills in which the same individuals have been studied over a period of two years.

You will recall that about two years ago, prior to your graduation, you took an objective test at your institution to help provide us with data which we needed for the first part of our study. Then, about a year later, you responded to our red-and-yellow follow-up card to verify your mailing address.

Because of the fine cooperation of people like yourself, our study has progressed very well according to schedule. However, in order to complete the project, it is necessary that we call upon you once more, even if you have neither entered the teaching profession nor had tests and measurements in your undergraduate or graduate work. Your participation at this time would involve about an hour of your time in filling out a brief questionnaire and taking an objective test. This could be done at your leisure. Any test scores or questionnaire responses, of course, would be held in the strictest confidence as research data. Please return the enclosed card and we will forward a set of materials to you within a few days after receiving it.

Without your cooperation and assistance at this final phase of the study, most of its value will be lost. In contributing some of your time to this project, you will help to increase understanding of the teacher-training process. We feel that this study can be of real importance and value to teachers throughout the country.

Sincerely yours,

Samuel T. Mayo, Director
Cooperative Research Project #5-0807
(formerly known as CRP #2221)

STM:acc
Enclosures

Appendix R

LOYOLA UNIVERSITY

*Lewis Towers * 820 North Michigan Avenue, Chicago 11, Illinois * WHitehall 4-0800*

May, 1966

Dear Colleague:

We appreciate your reply indicating your willingness to cooperate further in our measurement project. The questionnaire and test booklet are therefore enclosed as promised.

Our purpose in giving the test is to obtain a true picture of what you now know about testing, measurement, and evaluation. We would suggest that you answer the items on the test as spontaneously as possible, giving your first impression, even if some of the material seems unfamiliar. Two different kinds of objective test items comprise the test. They are the multiple-choice and the key-list types. It is essential that you follow the directions carefully as you go from a set of one type of item to another set.

For purposes of future mailing, please keep us informed of any changes in your address.

Again, you can be assured that all responses will be held in the strictest confidence as research data. On behalf of the project staff and the teachers who will benefit from this research, let me take this opportunity to thank you for your assistance.

Sincerely yours,

Samuel T. Mayo, Director
Cooperative Research Project #5-0807
(formerly CRP #2221)

STM: acc
Enclosures

Appendix S
FOLLOW-UP QUESTIONNAIRE AND TEST - FORM A

Department of Education - Loyola University, Chicago

POSTGRADUATION QUESTIONNAIRE

1. Name ____________________

2. Permanent Mailing Address: ____________________

____

3. Teaching Experience (Indicate your response by placing an "X" in the appropriate box.)

(1) ☐ Have taught continuously for two years

(2) ☐ Have taught continuously for one year

(3) ☐ Did not enter teaching

4. If you did not respond to box (1) above, briefly describe the job title or the nature of your work experience other than teaching and/or graduate study.

5. If you taught during the last two years, have you had any in-service training (such as lectures, non-credit workshops, or non-credit seminars)?

(1) ☐ Yes (2) ☐ No

6. If you answered "Yes" to question 5, please describe the content and the amount of time involved in the in-service programs in which you participated.

7. Graduate Study (Check one box.)

(1) ☐ Have taken graduate courses for credit.

(2) ☐ Have not taken any graduate courses for credit.

8. If you responded to (1) in question 7 above, please indicate the type of coursework taken or currently being taken and the number of quarter or semester hours.

Title of Course	No. Credit Hours Qtr.	No. Credit Hours Sem.

Now, please turn page and take the test.